COK-H01161-67-P014418 67

THE AMERICAN ECONOMIC REVIEW

SPINE — GREEN

391
·001

CALCUTTA UNIVERSITY CENTRAL LIBRARY

March–1974
JUN
VOLUME LXIV, NUMBER 1–3
64

GEORGE H. BORTS
Managing Editor

WILMA ST. JOHN
istant Editor

Board of Editors
BELA BALASSA
ROBERT E. BALDWIN
MARTIN S. FELDSTEIN
ANN F. FRIEDLAENDER
STEPHEN M. GOLDFELD
ROBERT E. HALL
BENT HANSEN
JACK HIRSHLEIFER
LEONARD A. RAPPING
SHERWIN ROSEN
THOMAS R. SAVING
ANNA J. SCHWARTZ
JEROME STEIN
JOSEPH E. STIGLITZ
S. C. TSIANG
FINIS WELCH
ARINA V. N. WHITMAN

Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A $10 submission fee must accompany each manuscript. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

No responsibility for the views expressed by authors in this REVIEW is assumed by the editors or the publishers, The American Economic Association.

Copyright American Economic Association 1974.

Articles

Limited Knowledge and Economic Analysis *Kenneth J. Arrow* 1

Myopic Optimizing and Rules of Thumb in a Micro-Model of Industrial Growth *Richard H. Day, Samuel Morley, and Kenneth R. Smith* 11

International Capital Movements and the Patterns of Economic Growth *Yusuke Onitsuka* 24

The Demand and Supply of Durable Goods and Durability *Richard W. Parks* 37

Competitive Price Adjustments: An Empirical Study *B. T. McCallum* 56

An Economic Model of Production and Investment for Petroleum Reservoirs *Robert G. Kuller and Ronald G. Cummings* 66

Can Equalization of Opportunity Reduce Social Mobility? *John Conlisk* 80

Toward a Theory of Multiple Customs Unions *Sukesh K. Ghosh* 91

The Dynamics of Inflation in Latin America, 1950–1969 *Robert C. Vogel* 102

The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model *Richard Dusansky and Peter J. Kalman* 115

Accounting for the Growth of Output *Spencer Star* 123

The Stability Properties of Optimal Economic Policies *Stephen J. Turnovsky* 136

Product Liability: The Case of an Asset with Random Life *John P. Brown* 149

Shorter Papers

Capital Markets and the Dynamics of Growth *Mathew Shane* 162
Labor Supply Under Uncertainty and the Rate of Unemployment *Michael J. Hartley and Nagesh S. Revankar* 170
Default Risk, Homemade Leverage, and the Modigliani-Miller Theorem *David P. Baron* 176
Patent Life and R&D Rivalry *Morton I. Kamien and Nancy L. Schwartz* 183
Life Cycle Consumption and Labor Supply: An Explanation of the Relationship Between Income and Consumption Over the Life Cycle *James Heckman* 188
Money Illusion and the Aggregate Consumption Function: Note *George D. Craig* 195
Semivariance and Stochastic Dominance: A Comparison *R. Burr Porter* 200
The Precautionary Demand for Reserves *Ernst Baltensperger* 205
Risk Premiums and the Life Cycle Hypothesis *Roger Bowden* 211
Private and Social Rates of Return to Education of Academicians: Note *Lorenzo Figà-Talamanca* 217
Private and Social Rates of Return to Education of Academicians: Note *John A. Tomaske* 220
Housing Market Discrimination, Homeownership, and Savings Behavior:
  Comment *John F. McDonald* 225
  Reply *John F. Kain and John M. Quigley* 230
The Administered-Price Thesis Denied: Note *J. Fred Weston, Steven Lustgarten, and Nanci Grottke* 232
Job Quitting and Frictional Unemployment *J. Peter Mattila* 235
Soviet Postwar Economic Growth and Capital-Labor Substitution:
  Comment *T. Krishna Kumar and Ephraim Asher* 240
  Reply *Martin Weitzman* 243

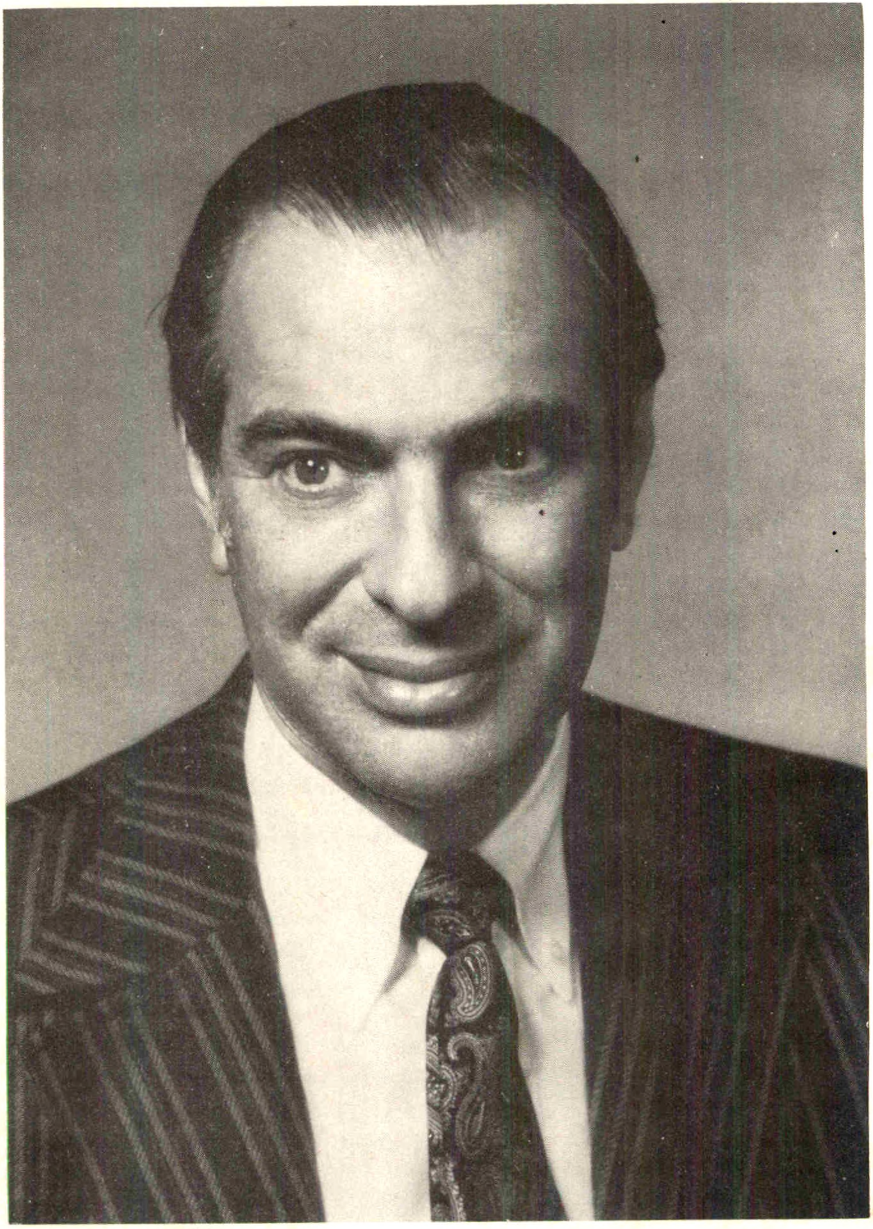
Notes 244

Announcements 247

P14418

Editor's Note: Under the terms of a resolution passed at the annual meeting of members of the American Economics Association on Dec. 29, 1973, $16.80 of the annual membership dues and subscription fee is considered as payment for the *American Economic Review* and *Journal of Economic Literature* the remainder is considered a contribution to exempt functions of the Association.

Number 75 of a series of photographs of past presidents of the Association



Kenneth J. Arrow

THE AMERICAN
ECONOMIC REVIEW

VOL. LXIV MAY, 1974 NUMBER 2

PAPERS AND PROCEEDINGS

OF THE

Eighty-sixth Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

New York, New York
December 28–30, 1973

Program Arranged by WALTER W. HELLER
Papers and Proceedings Edited by RENDIGS FELS
and MARK R. KILLINGSWORTH

Copyright American Economic Association, 1974

PRINTED BY GEORGE BANTA COMPANY, INC.

Publication Office: Menasha, Wisconsin 54952

Executive Office: American Economic Association
1313 21st Avenue South, Nashville, Tennessee 37212

Inquiries and other communications regarding membership, meetings, and the general affairs of the Association, as well as orders for the publications and permission to reprint, should be addressed to Rendigs Fels, Secretary of the American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212.

Changes of address should also be sent to the Secretary at the address given above. To be effective, notice must reach the office by the 1st of the month prior to the month of publication. The Association cannot be responsible for copies lost due to failure to report change of address in time.

Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, the American Economic Association.

Membership dues of the Association are $21 a year, $16.80 of each year's dues being in payment of a year's subscription to both the *American Economic Review* and the *Journal of Economic Literature*. The *American Economic Review* is published six times a year (March, May, June, September, October, and December) with four quarterly issues, the *Papers and Proceedings* of the annual meetings, the *Directory*, and supplements, and is sent to all members of the American Economic Association as one of the privileges of membership, as is the *Journal of Economic Literature* (four quarterly issues per year). Subscriptions by nonmembers (libraries, institutions, etc.) are $31.50 annually. Only subscriptions to both publications will be accepted. Single copies of the *Review* and *Journal* are $4.50 each, May *Papers and Proceedings*, $8.50. Each order for copies of the publications must include a $.50 per order service charge. Orders should be sent to the Executive Office.

Price $8.50

CONTENTS

Editors' Introduction . . . . . . *Rendigs Fels and Mark R. Killingsworth* vii

PAPERS

Richard T. Ely Lecture

The Economics of Resources or the Resources of Economics . . . . . . *Robert M. Solow* 1

The World Dynamics of Economic Growth

The Economics of the Steady State . . . . . . *Herman E. Daly* 15
Resources as a Constraint on Growth . . . . . . *William D. Nordhaus* 22
A Growth Model with Population Endogenous . . . . . . *Richard F. Kosobud and William D. O'Neill* 27

Population and Economic Policy

Resource and Amenity Implications of Population Changes . . . . . . *Ronald G. Ridker* 33
The Role of Population in Models of Economic Growth . . . . . . *Allen C. Kelley* 39
The Impact of Population Growth on Employment . . *Alan Sweezy and Aaron Owens* 45

International Energy Supply

The Political Economy of World Petroleum . . . . . . *James W. McKie* 51
Politics, Economics, and World Oil . . . . . . *M. A. Adelman* 58

Lessons From and For the Current Economic Expansion

Three Lessons . . . . . . *Paul W. McCracken* 68
A Strange Expansion . . . . . . *William J. Fellner* 71
Lessons from the Current Economic Expansion . . . . . . *Paul A. Samuelson* 75
A Plethora of Lessons from the Recent Expansion . . . . . . *Otto Eckstein* 78

Two Years of Wage-Price Controls

The Legacy of Phase II Price Controls . . . . . . *Robert F. Lanzillotti and Blaine Roberts* 82
Wages and the Pay Board . . . . . . *Daniel J. B. Mitchell and Arnold R. Weber* 88
The Inflation Problem during Phase III . . . . . . *Barry Bosworth* 93
Price Controls in 1973: Strategies and Problems . . . . . . *Mary T. Hamilton* 100

Monetary Policy and Sectoral Credit Flows

Alternatives to Monetary Policy . . . . . . *James S. Duesenberry* 105
Alternative Monetary-Fiscal Policies and Sectoral Credit Flows *Andrew F. Brimmer* 112

Major Economic Problems of the 1970's

The Clouded Crystal Ball . . . . . . *Robert L. Heilbroner* 121
Some Conjectures on Policy Problems of the 1970's . . . . . . *Kermit Gordon* 125
Economists as Servants of Power . . . . . . *Samuel Bowles* 129
Some Secular Changes in Business Cycles . . . . . . *Geoffrey H. Moore* 133
Policy Issues in the International Economy of the 1970's . . . . . . *H. S. Houthakker* 138

Concepts of Distributional Equity

Some Reasons for the Maximin Criterion . . . . . . *John Rawls* 141
Achieving the Just Polity . . . . . . *Dennis C. Mueller* 147
Political Constraints on Contractual Redistribution . . . . . . *James M. Buchanan and Winston C. Bush* 153
Discussion . . . . . . *Alvin K. Klevorick* 158

New Dimensions of Economic Inequality

The Concentration of Personal Wealth, 1922–1969 . . . . . . *James D. Smith and Stephen D. Franklin* 162

Who Paid the Taxes in 1966? .... *Benjamin A. Okner and Joseph A. Pechman* 168
Income Conditioning in a System of Transfer Programs .... *Barry L. Friedman and Leonard J. Hausman* 175

The Great Society Versus the New Federalism
Sorting Out the Social Grant Programs: An Economist's Criteria .... *Charles L. Schultze* 181
Cash Versus In-Kind Transfers .... *Lester C. Thurow* 190

The Fight Against Poverty
Earnings Capacity and the Target Efficiency of Alternative Transfer Programs .... *Irwin Garfinkel and Robert Haveman* 196
Conflicting Objectives in Income Maintenance Programs .... *Robert I. Lerman and Alair A. Townsend* 205

The Property Tax: Progressive or Regressive?
A New View of Property Tax Incidence .... *Henry Aaron* 212
Is a Property Tax on Housing Regressive? .... *Richard A. Musgrave* 222
Discussion .... *Harvey E. Brazer* 230
.... *Dick Netzer* 231
.... *Ann F. Friedlaender* 232
.... *Earl R. Rolph* 233
.... *George E. Peterson* 234

Women at Work: At Home and In the Market
Recent Trends and Long-Run Prospects for Female Earnings .... *Victor R. Fuchs* 236
Education and Home Production .... *Arleen Leibowitz* 243
The Effects of Government Expenditures and Taxes on Female Labor .... *Michael J. Boskin* 251

Production and the Aggregate Demand Function for Labor
The Elasticity of Substitution and Cyclical Behavior of Productivity, Wages, and Labor's Share .... *Thomas J. Sargent and Neil Wallace* 257
A Disequilibrium Model of Demand for Factors of Production .... *Sherwin Rosen and M. Ishaq Nadiri* 264

The Political Economy of Multinational Enterprises
Multinational Corporations and the International Adjustment Process .... *G. C. Hufbauer* 271
Competition Policy toward Multinational Corporations .... *Raymond Vernon* 276
Problems and Policies Affecting Labor's Interests .... *George H. Hildebrand* 283

Trends, Prospects and Policies for East-West Trade
Prospects and Policy for U.S.-Soviet Trade .... *Gregory Grossman* 289
Sino-American Trade Prospects and Policy .... *Alexander Eckstein and Bruce Reynolds* 294
Prospects for U.S.-East European Trade .... *Alan A. Brown, Paul Marer, and Egon Neuberger* 300

Perspectives on Twentieth Century Economic Theory
Capital Controversies: Ancient and Modern .... *John Hicks* 307

Is Economic Theory With It?
On the Relevance of the New Economics of the Family .... *Gary S. Becker* 317
Economic Theory and Its Discontents .... *Vernon L. Smith* 320

The State of Theory . . . . *Harry G. Johnson* 323
Is the Theory of Competitive Equilibrium With It? . . . . *Tjalling C. Koopmans* 325

The Theory of Policy
On Stabilization Policy: Goals and Models . . . . *Thomas Muench and Neil Wallace* 330
On Optimal Taxation with Costly Administration . . . . *Walter Perrin Heller and Karl Shell* 338

Allocating Resources for Policy Research
How Can Experiments Be More Useful? . . . . *Alice M. Rivlin* 346
The Use of Surveys for Policy Research . . . . *F. Thomas Juster* 355

Intercity Passenger Transportation
Airline Capacity Limitation Controls: Public Vice or Public Virtue? . . . . *George Eads* 365
Amtrak in Perspective: Where Goest the Pointless Arrow? . . . . *Edwin P. Patton* 372
The Wheel: Should We Reinvent It? . . . . *Hugh S. Norton* 378

The Economics of the First Amendment
The Market for Goods and the Market for Ideas . . . . *R. H. Coase* 384
Regulation and Diversity in Communications Media . . . . *William F. Baxter* 392
Discussion . . . . *Bruce M. Owen* 400

Economic Education
Developing Independent Problem-Solving Ability in Elementary Economics . . . . *Rendigs Fels* 403
Computer-Aided Instruction for Large Elementary Courses . . . . *Bernard H. Booms and D. Lynne Kaltreider* 408
Introducing Freshmen to the Social System . . . *Kenneth Boulding and Elise Boulding* 414

Endowment Income, Capital Gains and Inflation Accounting
The Investment Income Formula of the American Economic Association . . . . *Donald A. Nichols* 420
What is Permanent Endowment Income? . . . . *James Tobin* 427
A Plan for the Definition of Endowment Income . . . . *James M. Litvack, Burton G. Malkiel, and Richard E. Quandt* 433
Discussion . . . . *Robert Eisner* 438
. . . . *Stanley W. Black* 441

PROCEEDINGS

John Bates Clark Award 444
Minutes of the Annual Meeting 445
Minutes of the Executive Committee Meetings 451
Reports
Secretary . . . . *Rendigs Fels* 458
Treasurer . . . . *Rendigs Fels* 463
Finance Committee . . . . *Beryl W. Sprinkel* 465
Auditors . . . . *Arthur Andersen & Co.* 468
Managing Editor, American Economic Review . . . . *George H. Borts* 476
Managing Editor, Journal of Economic Literature . . . . *Mark Perlman* 483
Committee on Hiring Practices . . . . *F. Ray Marshall* 488
Committee on Economic Education . . . . *G. L. Bach* 512
Census Advisory Committee . . . . *Gary Fromm* 514
Representative to the International Economic Association . . . . *Fritz Machlup* 516
Representative to the National Bureau of Economic Research . . . . *Willard L. Thorp* 518
Committee on the Status of Women in the Economics Profession . *Carolyn Shaw Bell* 519

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. Persons of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

Editors' Introduction

THIS volume contains the papers and proceedings of the eighty-sixth annual meeting of the American Economic Association. "Proceedings" concern the business activities of the Association—presentation of honors and awards of the Association, the annual membership meeting, the twice yearly meetings of the Executive Committee, and reports of various Association officers and committees—and, as with the "Notes" section in each issue of the *American Economic Review*, are published in the hope that members will be informed of, and will participate in, the Association's affairs.

The "Papers," which constitute the greater part of this volume, are roughly equivalent to two regular issues of the *American Economic Review*, but are published under quite different procedures. An explanation of current Association policies concerning papers for the annual meetings may therefore be of some interest. For further elaboration on some of these policies, see the correspondence in the Union for Radical Political Economics (*URPE*) *Newsletter*, Oct. 1973, *5*, 28–36.

Early in the year the Association's President-elect (in 1973 Walter W. Heller, in 1974 R. A. Gordon) sets in motion the process which culminates in the December presentation and May publication of the papers. After consultation and comments, both volunteered and solicited,[1] from a wide range of persons, the President-elect decides on the topics of sessions (roughly twenty) at which papers will be presented, sets limits on the length of papers and invites persons to organize these sessions. Each session organizer in turn invites, again after consultation, several persons (usually two or three in number) to give papers on the topic of the session in question, and asks others to give comments on the papers. The papers, or rough drafts, are supposed to be sent to the editors of the *Papers and Proceedings* by November 15; we then read them and return them to their authors with comments and suggestions. (Students harried by term-paper deadlines may be interested to learn that in 1973 only seven of the sixty-seven papers presented here were in our hands by November 15 and that fewer than one-third had arrived by December 25!) In late December the papers and comments are read at the Association's meetings; we ask the authors to send us the final versions of their papers by January 2. After giving the final versions further scrutiny, we then in almost all cases send them to the printer to be published. (We do not generally publish discussants' comments, however, unless there is some special reason to do so.)

Thus, the rules under which these papers are published are quite different from those concerning articles appearing in regular issues of the *Review*. The papers published here are invited, not contributed; except in unusual circumstances they must be less than 4,000 (and sometimes 3,000 or even 2,000) words in length; each author is asked to avoid highly technical discussion and instead to contribute a paper which will be intelligible to economists not specializing in the subject it discusses. While we do edit (sometimes extensively) most papers to improve content and style and to satisfy space constraints, we do not subject papers to any refereeing process, and publication of any paper which we receive prior to the final printer's deadline and which satisfies (or can be cut to satisfy) space requirements is virtually guaranteed.[2]

---

[1] Suggestions are always welcome. See *Amer. Econ. Rev., 63*, Dec. 1973, p. 1062.

These policies are advantageous in a number of respects. The volume of papers can be published without the long delays inherent in publication by regular journals, and it consists of relatively short papers which concern a wide variety of subjects and can be understood by nonspecialists. Authors get a chance to report briefly on research just completed, to discuss topical subjects in an informal and engaging way, and to summarize longer forthcoming publications; readers get a chance to browse among a large number of articles which are outside their major areas of interest but which are not as specialized or ponderous as those sometimes found in regular journals. And while the papers are not scrutinized in the way regular journal articles are, they are (or are at least intended to be!) nontechnical, and thus do not need the thorough refereeing usual for articles submitted to journals.

Since the papers are invited rather than contributed, and since scheduling must be completed quickly, there is danger that session organizers may (if only for convenience) invite contributions from well-known persons at prestige institutions. In arranging the 1973 program, Heller urged session organizers to make every effort to seek out other persons likely to contribute papers of at least equal quality (see *URPE Newsletter*, p. 35)—a difficult, but not impossible, request. Gordon, 1974 president-elect, has as an experiment decided to schedule a limited number of sessions at which contributed papers will be read; and for 1974, as in the past, the Econometric Society (whose meetings are held with the Association's) will schedule several sessions of contributed papers in economic theory and econometrics.

RENDIGS FELS
MARK R. KILLINGSWORTH

---

[2] We would refuse to publish such a paper only if we felt that it was utterly without merit; no paper has yet been rejected on these grounds. However, the Executive Committee has established (see below, minutes of the meeting of March 2, 1973) another ground for rejection: if a paper cannot be cut to meet space requirements, we ask the author either to authorize us to submit it to the managing editor of the *Review* to be considered for publication (subject to the usual refereeing process) in a regular issue of that journal, or to withdraw the paper and submit it elsewhere.

THE AMERICAN ECONOMIC REVIEW

June 1974

VOLUME LXIV, NUMBER 3

GEORGE H. BORTS
Managing Editor

WILMA ST. JOHN
Assistant Editor

Board of Editors

BELA BALASSA
ROBERT E. BALDWIN
MARTIN S. FELDSTEIN
ANN F. FRIEDLAENDER
STEPHEN M. GOLDFELD
ROBERT E. HALL
BENT HANSEN
JACK HIRSHLEIFER
LEONARD A. RAPPING
STEPHEN RESNICK
SHERWIN ROSEN
THOMAS R. SAVING
ANNA J. SCHWARTZ
JEROME STEIN
JOSEPH E. STIGLITZ
S. C. TSIANG
FINIS WELCH
MARINA V. N. WHITMAN

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A $10 submission fee must accompany each manuscript. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

• No responsibility for the views expressed by authors in this REVIEW is assumed by the editors or the publishers, The American Economic Association.

• Copyright American Economic Association 1974.

Articles

General Economic Equilibrium: Purpose, Analytic Techniques, Collective Choice *Kenneth J. Arrow* 253

Capital Importation and Sectoral Development: A Model Applied to Postwar Puerto Rico *John J. Mingo* 273

The Political Economy of the Rent-Seeking Society *Anne O. Krueger* 291

The Empirical Relationships Between the Dividend and Investment Decisions of Firms *Eugene F. Fama* 304

On the Impact of Uncertainty on the Value and Investment of the Neoclassical Firm *Guy V. G. Stevens* 319

Land Use and Zoning in an Urban Economy *William J. Stull* 337

The Neoclassical Production Function *Lowell Gallaway and Vishwa Shukla* 348

The Social Costs of Input-Market Distortions *Daniel Wisecarver* 359

Private Incentives and Public Information *John M. Marshall* 373

Testing for the Existence of a Consistent Aggregate Index of Labor Inputs *Ernst R. Berndt and Laurits R. Christensen* 391

The Stochastic Structure of the Velocity of Money *John P. Gould and Charles R. Nelson* 405

Shorter Papers

Faculty Salaries: Is There Discrimination by Sex, Race, and Discipline? *Nancy M. Gordon, Thomas E. Morton, and Ina C. Braden* 419

The Rationale of the Mean-Standard Deviation Analysis:

- Comment *Karl Borch* 428
- Comment *G. O. Bierwag* 431
- Comment *Haim Levy* 434
- Reply and Errata for Original Article *S. C. Tsiang* 442

Externalities as Commodities:

- Comment *F. Trenery Dolbear, Jr.* 451
- Comment *Talbot Page and John Ferejohn* 454
- Reply *Robert A. Meyer* 460

On Taxation and the Control of Externalities:

- Comment *T. H. Tietenberg* 462
- Comment *Earl A. Thompson and Ronald Batchelder* 467
- Reply *William J. Baumol* 472

Consumer Preferences and Gains from Trade *Yeong-Her Yeh* 473

A Note on the Price Adjustment Mechanism for Rental Housing *Lawrence B. Smith* 478

Phase II, Cost-Push Inflation, and Relative Wages *Michael L. Wachter* 482

Earnings, Employment, and Racial Discrimination: Additional Evidence *B. F. Kiker and W. Pierce Liles* 492

On Reanalyzing the Harris-Todaro Model: Policy Rankings in the Case of Sector-Specific Sticky Wages *Jagdish N. Bhagwati and T. N. Srinivasan* 502

The Nature of Quantities in Market Disequilibrium *Herschel I. Grossman* 509

The Nature of Quantities in Market Disequilibrium *E. C. H. Veendorp* 515

In Memoriam: LUDWIG VON MISES 518

Notes 519

Limited Knowledge and Economic Analysis

By Kenneth J. Arrow*

The content of presidential addresses to this Association provides a fine example of a random variable with a high variance. It might even be a good subject for econometric analysis; the variation might be explained in terms of the economic conditions of the moment, previous intellectual investments, or even, for boldly interdisciplinary analysis, the psychological states or class origins of the speakers or the audience. But no doubt captious theorists like myself will object that the endogenous variable is not cardinally measurable and probably not even ordinally measurable; tougher-minded econometricians will worry about collinearity in the predetermined variables; and practical-minded policy analysts will see no discernible effect on the gross national product, the price level, or the balance of payments through effects on either fiscal policy or the stock of money. The last group, the policy-oriented, are perhaps the least accurate; at least according to Keynes, the effect of ideas on policy is dominant, though the lag may be as variable as and a good deal longer than that of the stock of money on money gross national product.

It is now more fashionable than it used to be for statisticians to be told to take a good look at their data before fitting models. Taking presidential addresses as our data, we find most frequently a review of the speaker's main research concerns but also expressions of methodological or ethical concerns, historical surveys of varying degrees of erudition and humor, and, least frequently, new points of view on significant problems of economics.

I am taking a somewhat different tack today; it will be an expression of discontents and expectations. As I shall try to argue, the uncertainties about economics are rooted in our need for a better understanding of the economics of uncertainty; our lack of economic knowledge is, in good part, our difficulty in modelling the ignorance of the economic agent.

Critical aspects of this need for reorientation of theory have been recognized by many scholars in the last quarter-century and particularly in the last decade. I view my remarks today as a summary and perspective on a widely shared development of thinking.

The starting point of discussion must still be the much-abused neoclassical theory. No really cohesive alternative which aspires to the same level of completeness exists. The neoclassical model is founded on two concepts, which are considerably different in nature. One is the notion of the individual economic agent, whose behavior is governed by a criterion of optimization under constraints which are partly peculiar to the agent, such as production functions, and partly terms of trade with the economic system as a whole.

---

* Presidential address delivered at the eighty-sixth meeting of the American Economic Association, New York, New York, December 29, 1973. The views expressed above are a personal synthesis of a widespread viewpoint which has appeared in many different contexts in the published work of and in some cases personal discussions with G. Akerlof, A. Alchian, G. Calabresi, R. H. Coase, H. Demsetz, P. A. Diamond, J. Green, F. H. Hahn, A. G. Hart, F. A. von Hayek, J. Hirshleifer, L. Hurwicz, C. Kaysen, F. H. Knight, A. Leijonhufvud, F. Machlup, J. Marschak, T. Marschak, J. Mirrlees, M. Nerlove, R. Radner, M. Rothschild, T. Schelling, H. A. Simon, A. M. Spence, and G. J. Stigler. For some additional aspects of the role of information problems in economics, see my lecture, *Information and Economic Behavior*, Stockholm 1974.

The other is the market; here, the aggregate of individual decisions is acknowledged, and the terms of trade adjusted until the decisions of the individuals are mutually consistent in the aggregate, i.e., supply equals demand.

The neoclassical theory, especially in its competitive form, can be and has been given a rich formal development. Parenthetically, one cause for the persistence of neoclassical theory in the face of its long line of critics is precisely that for some reason of mathematical structure, the neoclassical theory is highly manipulable and flexible; when faced with a specific issue, it can yield meaningful implications relatively easily. Although I intend to air complaints and desires for change today, I must express my unabashed admiration for the accomplishments of the neoclassical viewpoint. In its most formal statement, we simply use for analysis the equilibrium conditions of the individual agent and of the market, without inquiry as to how they come to hold. Yet even these statements turn out to yield revealing insights in the workings of resource allocation. Why have medical costs risen so rapidly relative to other prices since 1967? The upward shift in demand due to Medicare and Medicaid with a price-inelastic supply of physicians and hospitals provides a simple straightforward answer; I cannot really imagine how a Marxian or a neo-Ricardian would even approach the question, though I suppose they might dismiss it as unimportant. The explanation of environmental problems as due to the nonexistence of markets is similarly an insight of purely neoclassical origin. The now-demonstrated fact that flexible exchange rates are a feasible way of conducting international finance is a triumph of theoretical insights over practical men's convictions. More broadly, the shifts in long-run resource allocation as motivated by returns and, in particular, the absence of a secular trend in technological unemployment to the perpetual surprise of the layman fit in well with the neoclassical formulation but have no ready explanation in alternate models.

Of course, the implications of neoclassical theory have also been conspicuously falsified in important ways. Most notably, the recurrent periods of unemployment which have characterized the history of capitalism are scarcely compatible with a neoclassical model of market equilibrium. A post-Keynesian world in which unemployment is avoided or kept at tolerable levels by recurrent alterations in fiscal or monetary policy is no more explicable by neoclassical axioms, though the falsification is not as conspicuous.

Inequality in economic development among countries and among groups and regions within a country provides a second, somewhat complicated difficulty for neoclassical theory. A purely neoclassical answer would explain differences in per capita income by differences in physical and human assets per capita. This of course raises the further question, how this came to be, a question which would require a fully dynamic model to answer; but I think the more compelling problem is that the differences in income seem much too vast to be explained by factor differences. Indeed, in the presence of international trade and especially international capital movements, wage differences should be very strongly reduced compared with what would occur in autarchic states where domestic capital is the limiting factor. Hence, we come immediately to the explanation that there are differences in the production possibility sets of the different countries. This conclusion is a legitimate and important use of neoclassical analysis; but obviously it raises new questions, to which we will return.

I pass by the whole tangle of questions relating to the holding of money and the general level of prices. In its pure form, neoclassical theory is a theory of relative prices. Monetary theories vaguely related to it in spirit can be grafted on to it, but none have succeeded in achieving a genuine synthesis.

The two failures of the neoclassical explanatory mechanism reflect on its foundations in quite different ways. The existence of unemployment is clearly a direct contradiction to the notion of the smoothly clearing market. One must of course be aware that the official measure of unemployment is by no means a simple inequality between supply and demand; it aggregates a whole range of distinct markets, it does not separate out voluntary and involuntary unemployment according to the tests of economic theory, and it does not take account of unfilled jobs. I do not subscribe, I hasten to say, to the sometimes expressed view that all unemployment is essentially voluntary, an unwillingness to search or whatnot; indeed, the official measure may underestimate the degree of disequilibrium in the labor market, particularly with regard both to underutilization of advanced skills and discouraged job seekers. With all these qualifications, it is clear that statistical unemployment does correspond to a disequilibrium as that term is used in the basic neoclassical model; there are two individuals, identical in productive capacity and both willing to work at a given wage, but one is working at that wage and the other is not.

Differential levels of economic development, on the other hand, point to a difficulty with the other fundamental concept; the conditions of optimization. If countries differ in their production possibility sets, then firms, occupying similar economic positions, are facing different constraints on their optimization. This does not contradict the fundamental assumption of optimizing behavior, but it does raise severe questions about its interpretation. The simplest hypothesis is to take the technological conditions as data, possibly varying over time due to exogenous changes in scientific knowledge. But here we are asserting that two contemporaries have different access to productive knowledge. Clearly, we are saying something about the conditions of transmission of knowledge across national boundaries, and of course the same questions arise among firms or workers within a single economy. The constraints upon the firm's optimization begin to seem more like variables to be explained than like constants exogenously given.

Let me look now at the two basic concepts from the inside, from the point of view of our direct perceptions which motivate the modelling. The two are far from parallel. The optimization by individual agents has a sense of concreteness about it, for all the sophisticated mathematical ability with which we theorists endow the agents. They behave in ways whose logic we understand. They seek to achieve goals which are reasonable to postulate, and we can specify constraints which clearly are real. It can be and has been correctly objected that our models are too simple; we ignore other arguments in the utility function, power, status, social approval, or whatnot that also motivate individuals, and we ignore some constraints, capacity for calculation and social controls. But the model is comprehensible, and the motives and constraints we deal with are real and important.

The market, on the other hand, is a much more ethereal construct. Who exactly is it that is achieving the balancing of supply and demand? Where in fact is the information on bids and offers needed for equilibration actually collected and stored? Right from the beginning of neoclassical theory, the difficulty of explaining markets in terms of individual self-seeking behavior was perceived. Parenthetically, this is one example of the superiority of neoclassical analysis to its predecessors, despite the current fashion for exalting Ricardo over his successors; Ricardo implicitly equated supply and demand on all his markets without ever realizing the problematic nature of this process. Jevons felt obliged to enunciate explicitly a Law of Indifference, enforced by arbitrage; but this does not really meet the problem when the market is out of equilibrium, for the arbitrage might well not be feasible. Menger, at least according to Hayek and Streissler, concentrated on individual trades and ignored the market completely. It is Walras's auctioneer which has proved to have had the most enduring effect on subsequent theoretical development, and the stability theory which flows from that concept is still the subject of vigorous theoretical development, though very little empirical application. What is envisioned is a feedback mechanism in which errors in the price are successively corrected by reference to the disequilibria they generate. This view specifies and makes feasible the operations of the market. But on one hand the stability models are far from adequate representations even of the dynamics of the neoclassical models and, what may be connected, the results are by no means necessarily favorable to the stability of the adjustment process; and on the other hand, the motivations for the feedback to operate are obscure.

Let me be clear on one methodological point. The fact that our intuitive understanding, our *verstehen* as the German social methodologists call it, of the market as an institution is not entirely satisfactory does not mean that we should not use the perfect market as a model, at least pending further development. Certainly, as Popper and Friedman hold, the acceptability of a theory is to be judged by its ability to predict and understand phenomena. The theory of the perfect market is in an interesting way complementary to Keynesian theory. We have never been able to integrate Keynesian viewpoints into standard neoclassical theory, in terms of individual motivation, yet this theory, with its various modifications, has been a most serviceable tool of prediction and control. In fact, it is useful in domains where competitive theory fails and *vice versa.* Neither theory is good, however, at predicting dynamic processes, the short-run changes which are responses to disequilibria, and it is here that the pressure for a more satisfactory model arises.

Hold in abeyance for a moment our considerations about the market. Let us return to the optimization problem of the individual. One aspect on which we put a good deal of weight, particularly in our less formal discussions, is that a market system is informationally economical. That is, we tend to regard it as a virtue of the system that the individual agent need not know very much. Specifically, he is supposed to know the motivation and production conditions which define him, i.e., his utility function and production possibility set, together with the prices of the commodities he buys and sells. The economic system, taken as a whole, has vastly more in it than any one individual knows; it contains the utility functions and production possibilities of all individual agents. Indeed, the apparent modesty of the information needed is one of the most appealing aspects of the neoclassical model, both in the descriptive sense that the individual's decision problems appear manageable for him and for the economist studying him, and in the normative sense that the system permits its members to spend their time and effort at producing goods rather than in unnecessary duplication of information.

But clearly this simplification of the individual's decision making is made possible only because the markets have supplied the information economized on, in the form of prices. In equilibrium, at least, the system as a whole gives the impression of great economy in the handling of information, presumably because transmission of prices is in some significant sense much cheaper than transmission of the whole set of production possibilities and utility functions. It is this point which emerged in the great debate over the feasibility of socialism begun by Ludwig von Mises's attack and usually thought of as concluding with the work of Oskar Lange and Abba Lerner in the 1930's; though it should be added many of the essential points had already been made earlier by Vilfredo Pareto and Enrico Barone. What was argued, in effect, was that a socialist system could use the price system and therefore achieve whatever economies in information it does achieve; and if the equilibrium conditions are written out they do give the appearance of relative simplicity. But what was left obscure is a more definite measure of information and its costs, in terms of which it would be possible to assert the superiority of the price system over a centralized alternative. Though I feel that current work has brought about a considerable clarification, we still have no definite measure. Indeed, in some respects, more recent developments have made the answers less clear. Several writers, in both Western and socialist countries, have noted that alternative decentralized schemes exist where quantity messages rather than price messages are transmitted in the successive stages of approximation and that such schemes also have efficient equilibrium points. Indeed, with the development of mathematical programming and high-speed computers, the centralized alternative no longer appears preposterous. After all, it would appear that one could mimic the workings of a decentralized system by an appropriately chosen centralized algorithm. While there is more to the story than these few remarks, they do make the point that if we are going to take informational economy seriously, we have to add to our usual economic calculations an appropriate measure of the costs of information gathering and transmission.

But actually the comparisons between socialist and capitalist resource allocation systems have tended to overlook some of the most obvious facts while examining finer points closely. As we all know, both production and consumption decisions are in fact made with reference to the future as well as to the present. A rational production plan includes very importantly decisions or at least plans about the future; and similarly with consumption plans. Investment and savings are not only integral parts of our current decisions but in the long run shape the possibilities for further development. As we know, the formal neoclassical model can be extended to decisions over time by dating commodities and regarding the same commodity at different dates as different commodities. All previous conclusions follow; allocative efficiency, for example, is achieved with the same appearance of informational efficiency.

But of course there is a slight problem with this reasoning. The information about future commodities needed includes their prices. These prices must be those found on a suitable market, one in which future supply and future demand are equated. Unfortunately, no such markets exist. Even the futures markets in certain commodities, limited in extent as they are, do not in fact lead to balancing *all* future decisions. Rather they balance present commitments to the future; but it is understood by all parties that when the future becomes the present, there will be a spot market on which the futures commitments may be undone; and indeed those making no futures commitments at all can enter and know now that they will be able to enter.

Even as a graduate student, I was somewhat surprised at the emphasis on static allocative efficiency by market socialists, when the nonexistence of markets for future goods under capitalism seemed to me a much more obvious target.

However that may be, the nonexistence of these markets must be faced. Now in general equilibrium any part of the system affects every other part in at least two different ways. Thus, we may ask two questions about the nonexistence of futures goods markets: what are its implications for the rest of the system and what are the reasons for its nonexistence.

The implication first of all is that the information needed by the optimizer is not provided by an existing market. It will be provided by a market which will exist in the future, but that is a bit too late to help in decisions made today. Hence, the optimizer must replace the market commitment to buy or sell at given terms by expectations: expectations of prices and expectations of quantities to be bought or sold. But he cannot know the future. Hence, unless he deludes himself, he must know that both sets of expectations may be wrong. In short, the absence of the market implies that the optimizer faces a world of uncertainty.

The exact modelling of behavior under uncertainty is probably not crucial to the subsequent discussion; let us use the conventional expected-utility hypothesis. When there is uncertainty, risk aversion implies that steps will be taken to reduce risks. This partly affects decisions within the firm, such as the holding of inventories and preference for flexible capital equipment, and partly leads to new markets which will shift risks to those most able and willing to bear them, particularly through the equity market. The rich development of inventory theory and portfolio theory in the last twenty years or so reflects growing understanding of these matters.

But when we speak of expected utilities, we need to have some probabilities. Where do these come from? We may in the first instance regard them as subjective. But the economic agent observes his world and has the opportunity to learn from his experience, for there is a considerable degree of continuity. By Bayes' Theorem or perhaps psychological learning theory, the probabilities, say of future prices, will gradually adjust so as to conform to the facts. If indeed the economic world exhibited the same structure in some sense from period to period, and if everybody observed everything relevant, then the probabilities ascribed by different individuals to the same events might be expected gradually to converge to the correct values and therefore be the same for all. In fact, of course, the basic economic facts are changing, partly endogenously because of capital accumulation in its most general sense, partly exogenously with predictable and unpredictable changes in technology and tastes; equally if not more important, though, is the fact that the dispersion of information which is so economical implies that different economic agents do not have access to the same observations. Hence, it is reasonable to infer that they will never come into agreement as to probabilities of future prices.

A further implication is that the past influences the future. Jevons's well-known slogan, 'bygones are forever bygones," ceases to be fully accurate. The past is relevant because it contains information which changes the image of the future; the probabilities which govern future actions are modified by observations on the past. It follows that present decisions with implications for the future are functions of past values of variables as well as present values.

This point of view has been exploited in the econometric models which have used distributed lags in explaining investment decisions. What still needs to be exploited more, however, is that the inference to the future is necessarily uncertain, and the decisions made still exhibit risk aversion.

Expectations for the future are related to quantities as well as prices. The importance of quantity expectations has been stressed in macroeconomic models, even in such pre-Keynesian concepts as the acceleration principle, and most especially in relation to inventories. It sometimes is held that in a neoclassical world only prices matter; in the absence of prices, presumably they are replaced by price expectations. But that is not strictly true. Under constant returns, at least, quantity information for the individual firm is needed even when neoclassical assumptions are strictly fulfilled. Neoclassically founded investment theories usually predict capital-output ratios or capital-labor ratios; they still need output forecasts explicitly or implicitly. This gives considerable, perhaps major weight to past quantity information in predicting the future and therefore in guiding current investment decisions. It is perhaps along these lines that Keynesian theory, with its overwhelming emphasis on quantity changes as equilibrating variables, can be founded firmly on individual optimizing behavior.

I have referred to the fact that information is dispersed throughout the economy but have not suggested how. In the pure neoclassical model, each agent knows only his own production possibilities and his tastes, together with market information on the rest of the economy. In the world I have just sketched, however, any variables which improve his ability to predict the future have a very meaningful economic value to him. He will seek to acquire additional information. Such information is presumably costly; that is the basis for such great emphasis on the value of informational economy. But there is clearly a great incentive to acquire information of predictive value, and, as neoclassical theory would predict, there will be an incentive to produce such information. We have then an economic information industry: data assembly and analysis, business journalism, economic forecasting, with a longer-run perspective business education. Since information as a commodity does not satisfy all the neoclassical norms, it is not surprising that the government plays a large role in this process. Information-acquisition activities and information markets now appear on the economic landscape. Efficiency in the operation of firms ceases to be purely productive efficiency; it involves efficiency in prediction as well.

I would conjecture that the incomplete diffusion of information along the lines just sketched has a good deal to do with the operations of the securities markets and the decisions on corporate financing. The predominant role of internal financing and indeed the whole special importance of the managerial factor in corporate decision making are clearly connected with differential access to information about the firm.

You may have forgotten by now, but I earlier promised to consider not only the implications of but also the causes for the absence of markets for future goods. One might wonder why one should explain the absence of a phenomenon. Sherlock Holmes once maintained to the dimwitted local police inspector so typical of English detective stories that the significant question in the case at hand was the dog's barking at night. "But," said the inspector, "the dog didn't bark." "That," said Holmes, "is what is significant." So too is the absence of these markets significant for a full neoclassical theory. A truncated theory of temporary equilibrium in which markets for future goods are replaced by some form of expectations, themselves functions of current prices and quantities, has indeed been developed, though its empirical content is necessarily meager if the formation of expectations is left unanalyzed. But the true neoclassical spirit is being denied in such a model. Although we are not usually explicit about it, we really postulate that when a market could be created, it will be. I sometimes think that welfare economics ought to be considered an empirical discipline. Implicitly, if an opportunity for a Pareto improvement exists, then there will be an effort to achieve it though some social device or another. In our theories and to a considerable extent in practice, the cheapest way in many cases is the creation of a market; and markets do emerge. If a market is impractical for one or another of the reasons we usually call "market failure," then very likely some other social device will at least be tried: government intervention; codes of professional ethics; or economic organizations with some power intermediate between the competitive firm and the government.

Thus, the failure of markets for future goods must be regarded as an analytic problem as well as a presupposition. It seems to me there are two basic causal factors. One is that contracts are not enforceable without cost and forward contracts are more costly to enforce than contemporaneous contracts; the other is that because of the many uncertainties about the future, neither buyers nor sellers are willing to make commitments which completely define their future actions. Let me take these two points up in turn.

The ability to make enforceable contracts is a necessary but not sufficient condition for a market. However, there is no way to insure complete enforceability. An individual may make a contract which he cannot in fact fulfill. Penalties may indeed be imposed on failure to live up to one's agreement, but they are not a substitute for compliance from the viewpoint of the other party, and there is always a degree of cost in enforcing the penalties. The laws of bankruptcy are a social recognition that complete enforceability is impossible and that it is even socially desirable to set limits on the penalties for failure. However, when the exchange of values for values is simultaneous or nearly so, the contracts may almost be self-enforcing. If a good has been sold and not paid for, it can be recovered; if there is a continuing relation of buyer and seller, a failure to settle bills can be met by refusal to make further deliveries, in which case the loss is minimized. With contracts extending into the distant future, on the contrary, the possibility of failure to comply becomes greater, partly because the self-enforcement aspects become weaker, partly because unexpected changes may intervene to make even a sincerely intended compliance difficult or impossible.

The outstanding examples of forward contracts are credit instruments. The buyer, who is taking the risks of default, is motivated to protect himself by seeking more information about the seller. The lender wants to know the borrower's assets, the prospects for changes in them, possibly even what he is going to do with the money. This very individualized information-seeking relation is quite far from the arm's length impersonal model of a market. The so-called capital markets are in many structural aspects very different from our model markets. It is of course an empirical question how far their behavior departs from the model. But the recurrent theme of credit rationing and availability doctrines, the essential imperfections of the credit market which underlie monetarist theories of cyclical fluctuations suggest that the incomplete enforceability of credit contracts and the protective steps taken by lenders are significant factors in explaining the working of the market.

While the enforceability question explains why those forward contacts that are made do not constitute a perfect market, we need more to understand why even these are so limited in their coverage of future goods transactions. There are forward contracts in money, some commodities, real estate, but very little else. The explanation lies in uncertainties of both buyers and sellers about prices and quantities and about technology and tastes. Using uncertainties about prices and quantities as an explanation for market failure is a circular argument, though not necessarily a fallacy. That is, if all markets for future goods existed and cleared all transactions, then there would be no price-quantity uncertainties. But this much is true; if some markets for future goods do not exist, then the agents have uncertainties which are relevant to their behavior on markets for complementary or substitute goods. As Hicks showed a long time ago, complementarity and substitution can occur over time as well as simultaneously. If, as I will argue in a minute, uncertainty can tend to destroy markets, then we can conclude that the absence of some markets for future goods may cause others to fail.

To illustrate, the demand for capital goods at any point of time is dependent on the prices and sales of the product at future points of time. Therefore the demand for future capital goods will depend on expectations about the product at some still more removed time. If we assume only that we will not have markets for products at some distant point of time, then the resulting uncertainty will reflect itself in a failure of the market for capital goods in the nearer future, which will in turn create still further uncertainties.

Thus, if some markets for future goods are nonexistent, there will be uncertainties on the other markets; in addition, demand and supply conditions for the future are uncertain because of technological and taste shifts. Assume that both buyers and sellers are risk averters. Then without going into details it is reasonable to conclude that both demand and supply will have a downward bias as compared with the situation in which uncertainty is absent. A buyer will be unwilling to contract for purchase of a good if a superior or cheaper substitute may be available; and the seller will be unwilling to accept a price sufficiently low to be suitable to the buyer, particularly if he thereby precludes himself from a possible opportunity to shift his resources to other closely related goods. It would seem possible, at least, that there will be no price at which transactions in future goods will take place. From a theoretical viewpoint, one might say that the market is in a strange sort of equilibrium; there is some shadowy sort of price at which supply and demand are equated at zero. But this price is not performing much of a signalling function.

There is one ultra-neoclassical approach to the market treatment of uncertainties, in which I take some pride. That is the notion of a contingent market. Instead of letting uncertainty ruin existing markets, we can take it explicitly into account by buying and selling commitments to be carried out only if some uncertain event occurs. We could in principle imagine agreements to transact which will hold if and only a given conceivable technological innovation does not take place, with a second market for transactions valid if the innovation does take place. Then we can restore the possibility of markets.

Such contingent markets are not entirely unknown; insurance contracts are the purest example, and equity markets and cost-plus contracts provide more muddied illustrations. But they are relatively rare. Why this should be so follows again from the general problem of information costs and dispersal. If contracts are contingent on the occurrence of some event, then it must be verified whether or not the event occurred. But this is information, and as the example of a technological innovation suggests, it is information likely to be much more easily available to one party than to the other. Hence, the range of possible contingent contracts becomes limited to those for whom the events are easily verifiable for both parties. The implications of these limits are known in the insurance literature as adverse selection and moral hazard, and they are of immediate practical significance in such matters as health insurance. But more broadly, they so limit the scope of contingent markets in practice that, as argued before in connection with markets for future goods, they prevent the emergence of even technically possible markets because of the large unresolved uncertainties.

I hope enough has been said to indicate the widespread implications of costly, dispersed information for the process by which future-oriented economic decisions are made. Let me remark, briefly in view of the length of time I have already taken, that informational costs and values play a key role in modifying the structure even of contemporaneous transactions. The individual optimizing agent is supposed to know at least his technology or tastes and the prices he faces. We have already argued a good deal of uncertainty with respect to the future economic implications of present economic choices. But in addition there is the possibility that technological information, which would be useful to him, exists somewhere in the world but outside his firm. There are grounds for engaging in the active pursuit of information. We begin to enter the realm of diffusion of innovations, to which some sociologists as well as economists have contributed. The interesting points here are the biasses in the information channels, some of which, at least, can be explained in terms of differential costs of acquiring information. For example, the well-documented role of personal influence in accepting innovations can be interpreted as due to a perceived high reliability of such information; in economic terms, this means more information per unit of expenditure of time or money.

The terms of trade with the outside world should not be regarded as freely given to the firm. In a world with a large number of commodities, even knowing the prices of relevant commodities involves the costly acquisition of certain kinds of information. This remark has given rise to a large literature on search in recent years. One implication which has been only slightly explored is that the concept of the market begins to weaken, and Jevons's Law of Indifference becomes more of an equilibrium condition than a statement valid about a market even in disequilibrium. At a moment of time, prices of what would usually be thought of as the same commodity bought or sold by different firms can differ because buyers or sellers may not, in their ignorance and in the presence of costs of search, find it worthwhile to shop further. Obviously, the important application of this principle may be to the labor market. Clearly, there are important informational differences between the employees currently working for a firm and potential substitutes elsewhere, although these are interchangeable in pure neoclassical theory. Indeed, there are differences both in the information the firm possesses about its employees as compared with alternatives and the information which employees have about the economic opportunities and the specific production conditions of the firm as compared with outsiders. It appears that considerations of this type must play some role in understanding the continued possibility of unemployment and particularly the sluggish response of wages to market disequilibria.

I am far from exhausting the implications of an information-economical viewpoint for the economic world. I look forward to exciting developments in the next decade.

Myopic Optimizing and Rules of Thumb in a Micro-Model of Industrial Growth

By RICHARD H. DAY, SAMUEL MORLEY, AND KENNETH R. SMITH*

It has been argued by William Baumol and Richard Quandt (1964) and Day (1967) that rules of thumb can be efficient economic strategies when decision making is costly and when decision makers have imperfect information. The present paper augments this literature and analyzes an industrial growth model in which firms determine production and investment levels by solving a single period optimizing problem. The solution is carried out period after period rather than by making a complete lifetime plan involving forecasts of the future. We portray the firm as myopically groping toward an unknown equilibrium through successive one-period movements made by a simple linear programming allocation of the firm's cash budget.[1] Cash availability is the channel through which market feedback, operating through the demand function, modifies behavior. Our model can be thought of as a dynamic, rule of thumb, approximation to an intertemporally optimal investment and production path for the firm and industry.

It is shown that, under certain conditions of demand and cost, an industry whose firms use our myopic investment rules will converge asymptotically to a perpetually maintainable capital stock. This solution is the long-run equilibrium of the perfectly competitive industry in which price just covers total costs of the marginal firm. In other words, a short-sighted optimization based on complete ignorance of demand can lead to equilibrium in the sense of theories based on perfect knowledge and polyperiodic time horizons. This is only one of several possible outcomes, however, for under somewhat different market conditions fluctuations in investment and production levels will eventually occur, perhaps after a protracted period of growth. The model can also generate S-shaped industry growth paths, simultaneous saving and investment, and investment at less than the maximum possible rate (excess borrowing capacity). All of these are commonly observed patterns of industry behavior.

The possibility of industrial instability suggests the need for a risk-avoiding rule at the firm level. The rule introduced is the safety-first principle (see Andrew Roy, Day, Dennis Aigner and Smith), a device that is easily seen to reduce the likelihood of unstable oscillations about industrial equilibrium.

I. The Firm

A firm during period t is assumed to produce a single commodity in amount x_t at constant unit cost, c, using the services of a single durable capital good which itself can be purchased at the beginning of each production period at constant unit cost, q. The firm has inherited some stock of durable capital, $(1-\delta)k_{t-1}$, where δ is a constant depreciation rate. It also has a stock of financial wealth, purchasing power or "working capital," m_t, which it can divide between current productions costs, cx_t, internal investment, qy_t, and external investment or "banking," s_t. Central to our model is the notion that there is a time lag between the incurrence of costs and sales receipts. Hence cash or working capital must be invested in production and enters as a constraint on the output of the firm. The output x_t is produced during period t and sold in the market at the end of the period at a price not known at the time production and investment are determined. We first consider rules that ignore the possibility of having inadequate working capital in the next period. Later, in Section III, the investment and production rules are modified to take uncertainty into account.

---

* University of Wisconsin, Madison. We were supported in part by the National Science Foundation, Rice University, and the University of California, San Diego, respectively.

[1] Static theories of the firm that explicitly incorporate the capital budget go back at least to Sune Carlson, whose study was published first in 1938. More recent programming formulations are typified by, for example, Baumol and Quandt (1964). Our approach imbeds the capital budgeting problem in a dynamic setting in a way closely related to the work of John Williams. The result is a model similar to the generalized cobweb model of Day and E. Herbert Tinney (1969) and the methods employed by them are the basis of the analysis presented here. We incorporate results reported in an earlier paper by Day and Richard Benson.

If capital is measured in units of output so that one unit of capacity is required for one unit of output, then the capital constraint is

$$x_t \leqq (1-\delta)k_{t-1} + y_t \tag{1}$$

The cash or working capital constraint is

$$cx_t + qy_t + s_t - b_t \leqq m_t \tag{2}$$

where borrowing b_t is constrained to be less than or equal to some proportion of the available working capital:[2]

$$b_t \leqq \theta m_t \tag{3}$$

Borrowing is assumed to be for one period and m_t is assumed to have been adjusted for the retirement of the preceding period's debt. Finally, the nonnegativity constraints are

$$x_t,\ y_t,\ s_t,\ b_t \geqq 0 \tag{4}$$

Assume for the moment that expected price p_t' is given and that the firm borrows at rate r, which may be greater than i, the rate of return on external financial investment. Then the firm is assumed to choose its production, investment, and borrowing levels so as to maximize the one-period expected profit

$$\pi_t' = (p_t' - c)x_t - \delta q y_t + i s_t - r b_t \tag{5}$$

subject to (1)–(4).

The procedure is illustrated in Figure 1. MM' is the internal cash constraint, i.e., working capital without borrowing. It has slope $-c/q$. NN' is the working capital constraint *with* borrowing up to the limit. KK' is the physical capital constraint. It may be binding as shown in (a), (b), and (d), or slack as shown in (c). RR' identifies a kinked isoprofit curve. Below the internal budget constraint it has slope $[p'-c(1+i)]/q(\delta+i)$. Above it, when borrowing occurs, it has slope $[p'-c(1+r)]/(\delta+r)q$. The change in slope reflects the higher opportunity costs of obtaining funds from outside sources when $r>i$. The set of feasible investment, production choices is the shade area. Each case (a)–(d) is associated with a particular rule. A fifth rule occurs when the isoprofit lines are negatively sloped. In this case all working capital is invested externally because production is unprofitable.

The five rules of thumb associated with these cases may be enumerated as follows:

I. *Produce, Invest, and Borrow* if:

A. $p_t' \geqq c(1+r) + q(\delta + r)$

or

$$\frac{p' - c(1+r)}{q} - \delta \geqq r$$

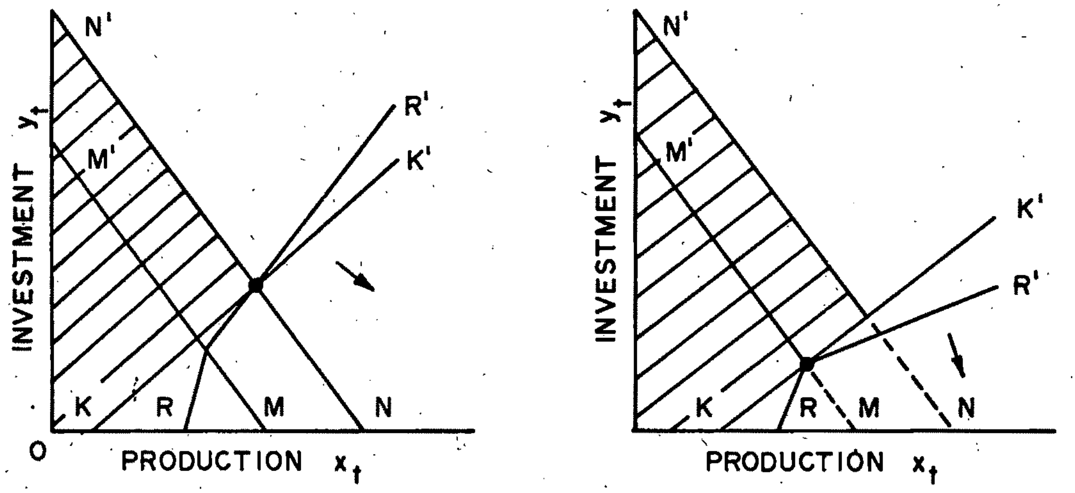
---

[2] In reality, credit rationing policies are reflected in the parameter θ which can be manipulated by banks in response to changing aggregate economic conditions, reserve banking policies and characteristics of specific industries. A useful extension of the present study would allow for many industrial sectors and a banking sector which adaptively modifies the θ parameter for each sector to bring about financial and product market equilibrium.

and

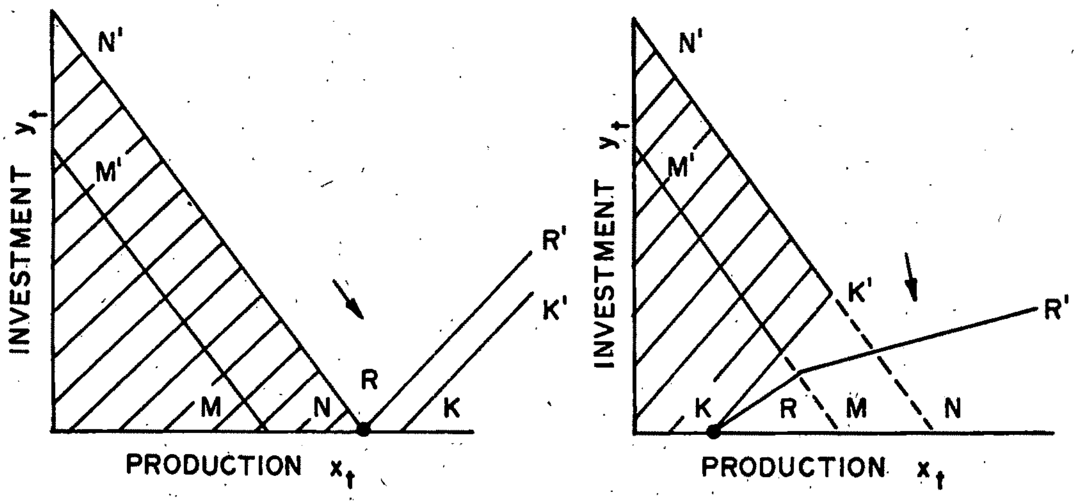
$$\text{B.}\quad \frac{(1+\theta)m_t}{c} \geqq (1-\delta)k_{t-1}$$

Condition A asserts that the firm should borrow and invest whenever price is greater than fixed plus variable cost or equivalently when the net internal rate of return is greater than the borrowing rate. Condition B asserts that available funds



PRODUCE, BORROW, INVEST
(a)
I

PRODUCE AND INVEST
(b)
II



PRODUCE AND BORROW
(c)
III

PRODUCE AND SAVE
(d)
IV

FIGURE 1. DECISION RULES FOR THE FIRM

exceed the amount needed to fully employ inherited capacity. *This is the maximum growth rate rule.*

II. *Produce and Invest, Finance Internally* if:

A. $$c(1+i)+q(\delta+i) \leqq p' < c(1+r)+q(\delta+r)$$

or equivalently

$$i \leqq \frac{p_t' - c(1+i)}{q} - \delta < \frac{p_t' - c(1+r)}{q}$$

B. $m_t/c \geqq (1-\delta)k_{t-1}$

Condition A asserts that the net internal rate of return on investment is greater than the lending rate but less than the borrowing rate. It is therefore profitable to invest from retained earnings only. Condition B asserts that internal funds must be sufficient to allow investment. This rule can, of course, only occur when $r>i$.

III. *Produce and Borrow for Production Only* if:

A. $p_t' > c(1+r)$

or equivalently

$$\frac{p_t' - c}{c} > r$$

B. $$\frac{(1+\theta)m_t}{c} \leqq (1-\delta)k_{t-1}$$

Condition A asserts that price is greater than variable cost or that the internal rate of return on funds invested in production is greater than the cost of borrowing. Condition B asserts that there is excess capacity relative to the supply of working capital including borrowing.

IV. *Produce and Bank, Do Not Invest* if:

$$c(1+i) < p_t' < c(1+i)+q(\delta+i)$$

or equivalently

$$\frac{p_t' - c}{c} > i; \qquad \frac{p_t' - c(1+i)}{q} - \delta < i$$

The condition asserts that price covers variable but not fixed cost. Production using inherited capacity and working capital is profitable, but capacity additions are not. If more capacity is available than can be used given available working funds the excess should be allowed to stand idle and to depreciate.

V. *Do Not Produce or Invest, Bank Only* if:

$$p_t' < c(1+i) \quad \text{or} \quad \frac{p_t' - c}{c} < i$$

In this case retained earnings cannot be invested profitably in the firm and should therefore be invested externally or banked.

In short, the firm produces whenever expected price, p_t', is greater than variable cost, $c+ic$, and invests whenever price is greater than variable plus capital costs, $c+ic+q[\delta+i]$. The firm borrows if its cash constraint is binding and its expected revenue covers total cost evaluated at the borrowing rate r. We will make use of these facts in determining the possible time paths of the complete model. Before doing so, we must specify a price expectation model, a dividend policy, and define the source of working funds. We consider these in turn.

We specify initially the simplest possible, naive price forecast:

$$p_t' = p_{t-1} \tag{6}$$

Next, we assume that firms have a target dividend policy of paying out interest on stockholders' equity, defined as the value of the capital stock plus the direct cost of production. We recall that because of the production lag, the latter is a part of the owner's total equity tied up in the industry. We also assume that this is supplemented by a marginal dividend increase or decrease β based on supernormal profits, and that if losses are great enough, dividends are reduced to zero.[3] If d_t are dividends paid out at the end of the period these assumptions mean that:

$$(7)\quad d_t = \max\{0, i[qk_t + cx_t] + \beta[p_t - (1+i)c - (\delta+i)q]x_t\}$$

Alternatively we may write dividends as

$$(7')\quad d_t = \max\{0, i[m_t + q(1-\delta)k_{t-1}] + \beta[\cdot]\}$$

a form which shows that working capital must include a "sinking fund" for plant and equipment. Next, we assume that the source of internal working funds is past sales and past external investment less debt repayments. Given the assumed dividend policy, internal working funds available at the beginning of period t+1 are sales revenue plus "bank funds" less borrowing repayments and dividends:

$$(8)\quad m_{t+1} = p_t x_t + (1+i)s_t - (1+r)b_t - d_t$$

This is the fund that can be "put to work" either inside or outside the firm during the coming period. Finally, assuming constant depreciation δ, we have

$$(9)\quad k_t = (1-\delta)k_{t-1} + y_t$$

We now have a complete model of the firm. It is worth noting that it is a close analogue of business and farm "budgeting" practice and hence is a reasonably plausible representation of actual behavior.

II. The Industry

In this section we discuss the possible growth paths of an industry composed of firms using the plausible rules of thumb described above. We show that industry equilibrium is the long-run equilibrium of a perfectly competitive industry. Convergence to equilibrium may be oscillatory or asymptotic or the equilibrium may be unstable, depending on demand conditions.

Let $X_t^D = f[p_t, w_t]$ be an industry demand function, where X_t^D is the quantity demanded by consumers, and w_t is a list of exogenous variables representing the influence of all other prices and income. Let $\sum_i x_{it} = X_t$ be total industry output, summed over all the firms in the industry. Assume that the market clears at a common price for each firm. Then we may write

$$(10)\quad p_t = F(X_t, w_t)$$

We shall assume that this price-output function is bounded, continuous and downward sloping for all w_t.

The feedback effect of market demand on current decisions of the firm is represented explicitly by substituting (10) into (8), where dividends are given by (7). Working capital for the ith firm in year t is thus given by

$$(11)\quad m_{it} = F(X_{t-1}, w_{t-1})x_{i,t-1} + (1+i)s_{t-1} - (1+r)b_{i,t-1} - d_{i,t-1}$$

Output decisions by each firm, totaled for the industry, determine a market clearing price, which in turn determines expected future price and hence present output decisions.

Equations (7), (9), and (11) together with the decision rules I–V form a closed dynamic system. Given initial values for m_{i0} and k_{i0} each firm determines x_{i0}, y_{i0}, s_{i0}, and b_{i0}. Summing output over all firms, equation (10) gives base period price, equal to p_1' by (6). The capital constraint is $(1-\delta)(k_{i0}+y_{i0})$ and equation (11) gives m_{i1}. Then from rules I–V we obtain a solution $x_{i,1}$, $s_{i,1}$, $y_{i,1}$. Output can again be summed to get X_1 and the whole process repeated. In this way the behavior is generated over time of each firm in the industry and of the industry as a whole.

It is the behavior of the industry as a whole that we now wish to explore. The analysis is greatly simplified and nothing

---

[3] This dividend policy, while plausible, is not given a rational foundation in the present paper. However, it does display appropriate properties at equilibrium.

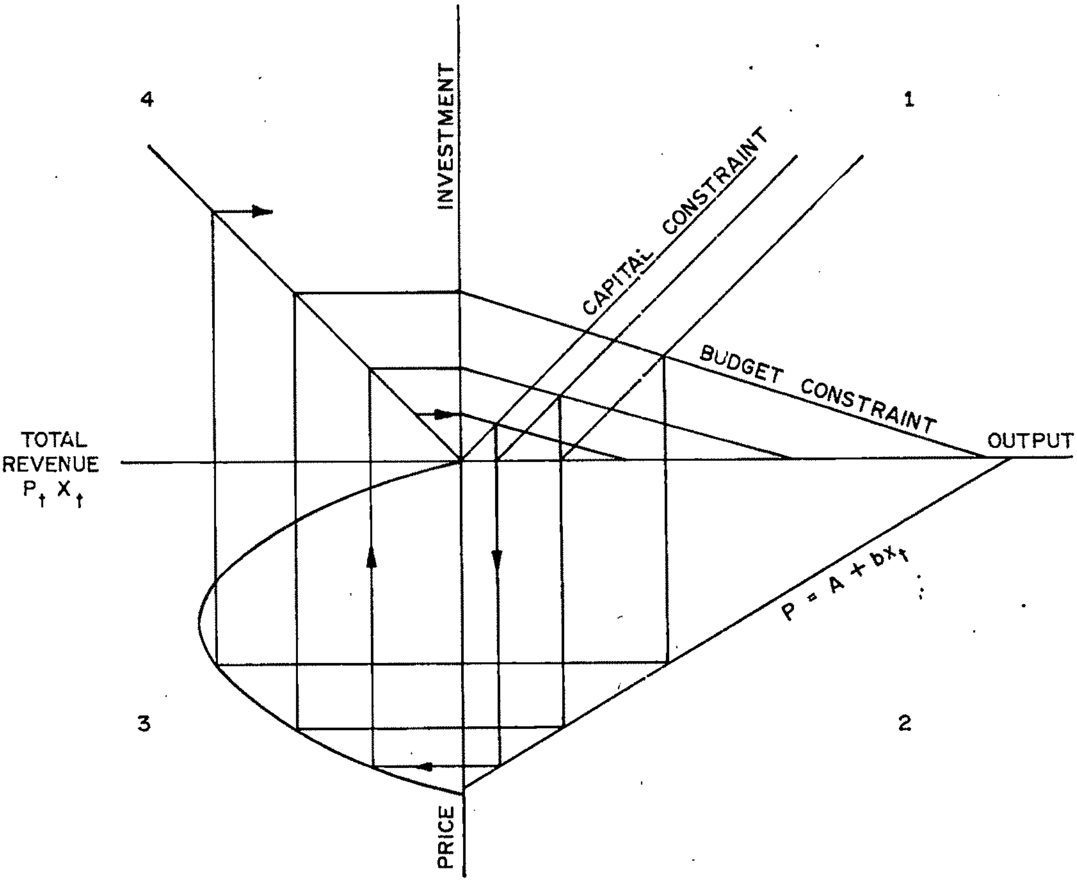


FIGURE 2. A GRAPHICAL CONSTRUCTION OF INDUSTRY BEHAVIOR

(*Note*: Depreciation, dividends, and borrowing assumed zero. The spiral moves counterclockwise. Note how the budget constraint shifts outward while the capacity constraint shifts upward.)

essential lost if we assume all firms have identical cost coefficients, price expectations, dividend policies and initial ratios of working funds to available capital (m_{i0}/k_{i0}). Then in each period all firms will face the same expected profits and the same ratio of working funds to capital stock. All firms will behave in exactly the same way and rules I–V can be used to describe aggregate behavior for the industry. Obviously the behavior of the industry will be identical to that of the individual firm.

The dynamics of industry behavior can be displayed graphically by temporarily making a few further simplifying assumptions. Assume no depreciation $(\delta=0)$; assume no dividends are paid so that the equation $d_t=0$ is substituted for equation (7); also let $q=1$ and assume $\theta=0$ so that no borrowing is allowed. Finally assume demand is stationary so that w_t is constant over time. Now consider Figure 2.

The upper right quadrant 1 represents the industry's feasible region for aggregate investment and production. In the lower right quadrant, the inverse industry demand curve is drawn so that for any given industry output the realized market price can be projected onto the left axis. This gives the expected price for the next period's decision. Actual total revenue received is plotted against price received in quadrant 3. Since for the sake of this illustration we assume dividends are zero, this is the maximum amount of funds available the next period and determines the intersection of the new budget constraint with the investment axis. Hence in quadrant 4 a 45° line enables us to project new working funds and in this way determines the feasible region for the next period. Neoclassical zero profit equilibrium exists at

$$P = c(1 + i) + q(\delta + i) = c + i(c + q)$$

(given $\delta = 0$), a horizontal line somewhere in quadrant 2.

The reader can experiment with demand curves of various shapes and cost coefficients at various levels to see how the growth path of the industry and its behavior around equilibrium depend on the elasticity of demand at equilibrium. We note that when demand is elastic, cash available to the industry increases as the industry grows so that the budget constraint shifts outward over time; but when demand is inelastic, the industry's working capital budget contracts when output expands. It is for this reason that converging, stable, and diverging cycles are all possible. Since the amount of investment depends on cash available, borrowing would speed up the convergence toward equilibrium, but it could also increase the amount by which equilibrium is overshot. Presumably, lenders could modify this risk by lowering θ, the borrowing ceiling. They might be expected to do this in industries with an historic tendency to cycle.

It is evident that depreciation, if allowed, would have the effect of moving the equilibrium price upward toward the elastic region of the demand curve and would have a stabilizing effect. Moreover, the tendency toward destabilizing overreaction might be moderated by changing the naive price forecasting rule to one based on a weighted average of several past prices. It can also be seen that when industry output is initially small the growth path will be S-shaped. This is because the ability to invest depends on retained earnings which accumulate most rapidly in the low output, high price portion of the industry demand curve. Borrowing would steepen this portion of the path.

The variety of possible outcomes of the industry adjustment process can be effectively summarized with reference to the industry demand curve and short- and long-run marginal cost curves as shown in Figure 3.

Given demand curve D_1, equilibrium for the industry is at E_1. At this point, price equals long-run marginal cost, $(\bar{p} = (1+i)c + (\delta+i)q)$, the familiar result of static, neoclassical economic theory. Also, equilibrium investment is the amount needed to maintain a constant, fully employed capital stock, $\bar{Y} = \delta\bar{X}$, and dividends are exactly equal to the normal return on funds invested in working and fixed capital. If the demand curve is D_2, long-run equilibrium with positive output is not possible, and if D_3 prevails, output even in the short run is uneconomic. None of these results is surprising, though of course here,

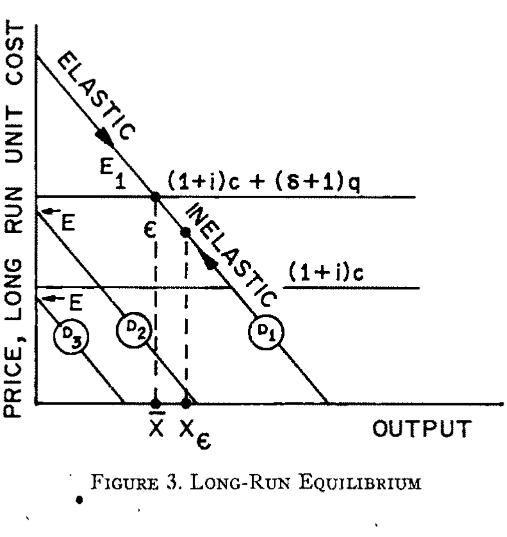


FIGURE 3. LONG-RUN EQUILIBRIUM

unlike the neoclassical theory, they are characteristics of myopic optimizing rules of thumb in equilibrium.

The most interesting results, which we already have been at pains to emphasize, involve the analysis of when this long-run equilibrium is brought about. The discussion of Figure 2 already suggests that these involve a variety of interesting cases. For example, suppose initial output is greater than $\overline{X}$. Price covers variable but not fixed cost. The investment rules tell us that firms will produce up to the smaller of their cash or capital constraints but they will not invest. With depreciation the capacity and the output of the industry will fall toward $\overline{X}$. The possibility of overshooting equilibrium from a path of initial growth, a possibility to which we have already alluded, is sure to occur if long-run equilibrium output is greater than the point where unit elasticity occurs in the market, i.e., when $\overline{X}>X_{\epsilon}$. If output is less than $\overline{X}$, each firm will invest up to the limit of its cash resources. This could easily increase production to above $\overline{X}$, resulting in overcapacity. Further investment is unprofitable and capacity is allowed to depreciate until it is less than or equal to $\overline{X}$. If depreciation has caused capacity to fall below equilibrium, investment will again be profitable, and the process will repeat itself. Thus some sort of oscillation around equilibrium will occur, the amplitude depending on the size of δ.

An extreme possibility is for overcapacity to be so great that industry output brings about a price below even variable cost. According to rule V, the industry would be abandoned even though the profitability of entry for each firm would quickly be reestablished so long as there is some output at which consumers are willing to pay the full cost of production. The irrationality of such an extreme oscillation would be a thought that would no doubt occur to decision makers within the industry. They would have some rule in addition to rules I–V to protect themselves from such a situation. We shall take up this question in Section III.

Before doing so we recognize that when demand is sufficiently small so that at no output level can price cover full cost including depreciation, the industry gradually declines, allowing its capacity to wear out without repair or replacement, but producing as much as it can until its working capital peters out altogether. When demand is still lower, such as in D_3, an immediate demise is justified and entry would never be invited.

III. Risk Aversion

In the above presentation we have seen that the industry adjustment process contains a variety of possible outcomes. In certain circumstances the use of a myopic optimizing rule of thumb leaves the firm satisfied with its decision in the sense that capacity was expanded at a maximum rate and further expansion is desirable. While price expectations may not be fulfilled, the use of a naive price forecast may nevertheless result in fulfillment of the firm's qualitative expectation that investment is profitable. In other circumstances, however, the firm's qualitative expectation will not be fulfilled. Investing in new capacity at the maximum rate results in overshooting the equilibrium and, thus, excess capacity. This possibility may lead the firm to more cautious behavior than that described in Section I.

In particular, the industry tendency to destabilizing over-reaction will be reduced by introducing risk aversion behavior on the part of the firm. We incorporate this idea here by introducing a safety constraint designed to prevent a disastrous drain on liquidity. It is derived from Roy's strict safety-first principle by assuming that the firm acts *as if* its total revenue were a random variable with standard deviation σ. It is not unreasonable to suppose that firms might act in this way for at least two reasons. In the first place, if all firms are investing, the firm will not know what their combined effect will be on total industry capacity. The change in capacity, combined with the industry elasticity of demand will determine a price decrease, given the demand function. Combined with this capacity effect is a second effect resulting from the growth over time in industry demand. There is no reason why this growth rate should be constant or that firms will know the rate. Whereas the capacity effect tends to decrease expected price, the growth effect tends to increase it. Explicitly considering the price as if it were a random variable influences the firm's behavior by affecting the internal working capital which the firm can plan to carry forward into the next period. If price were to be substantially less than expected, the firm could be left without sufficient internal funds to continue operations. The greater the chance that such an undesirable event will occur, the more risky it will be to follow the myopic rule.

Suppose the firm treats next period's anticipated working capital $\hat{m}_{t+1}$ as if it were a random variable with expected value:

$$E(m_{t+1}) = p_t' x_t + (1 + i)s_t - (1 + r)b_t - d_t'$$

where

$$\begin{aligned} d_t' = \max \{0,\ & i(qk_t + cx_t) \\ & + \beta[p_t' - (1 + i)c - (\delta + i)q]x_t\} \end{aligned}$$

and variance σ^2. It can be shown that:

$$P_r\{m_{t+1} \leqq M_D\} \leqq \frac{\sigma^2}{(E(m_{t+1}) - M_D)^2} \leqq \alpha$$

where M_D is a disaster level of working capital. The certainty equivalent is therefore

$$\begin{aligned} (12) \quad E(m_{t+1}) = p_t' x_t + (1 + i)s_t - (1 + r)b_t \\ - d_t' \geqq M_D + \frac{\sigma}{\sqrt{\alpha}} = \tilde{M} \end{aligned}$$

That is, the firm adds a margin of error to its estimate of future cash requirements to guard against unexpected "short-falls" in actual revenues, a fairly plausible rule of thumb.

Using the definitions of saving, borrowing, and dividends, and relation (2), equation (12) can be rewritten as

$$\begin{aligned} (13) \quad [(p_t' - (1 + i)c)(1 - \beta) + \beta q(\delta + i)]x_t \\ - q(1 + i)y_t \geqq \tilde{M} - m_t + iq(1 - \delta)k_t \end{aligned}$$

when the firm is banking, and as

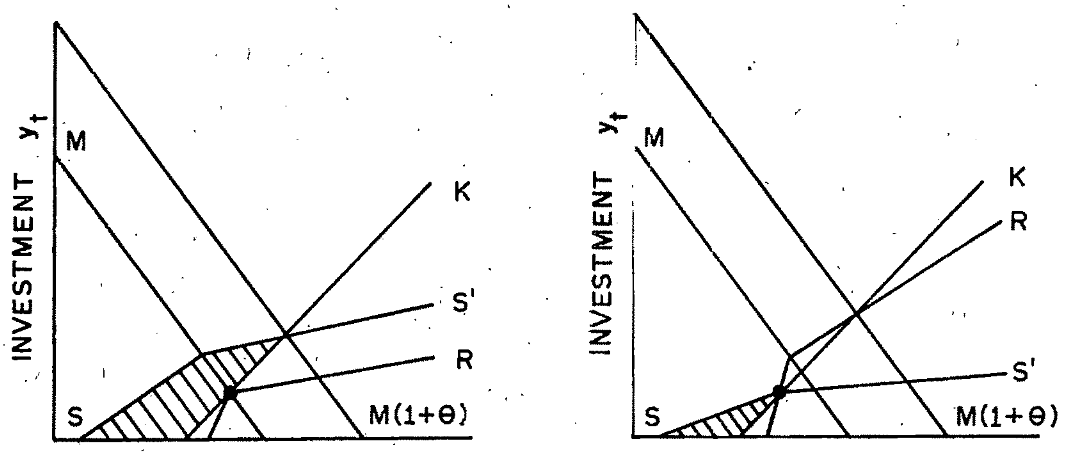
$$\begin{aligned} (14) \quad [p_t' - (1+i)c(1-\beta) + \beta q(\delta+i) - (r-i)c]x_t \\ -(1+r)qy_t \geqq \tilde{M} - (1+r-i)m_t + iq(1-\delta)k_{t-1} \end{aligned}$$

when the firm is borrowing.

The coefficient of x will be positive whenever production is profitable. It is the addition to end of period working capital per unit of output. The coefficient of y measures cash drain per unit of investment. The left-hand sides of (13) and (14) show the amount of working capital above present levels which must be generated through production. Generally one would expect the left-hand sides of (13) and (14) to be positive. That is, desired working capital plus the payment of dividends on capital should usually be greater than working capital on hand. If production is profitable, the greater output is, the greater expected working capital at the end of the period will be. It is therefore only where investment is taking place that the expected working capital constraint could be binding.

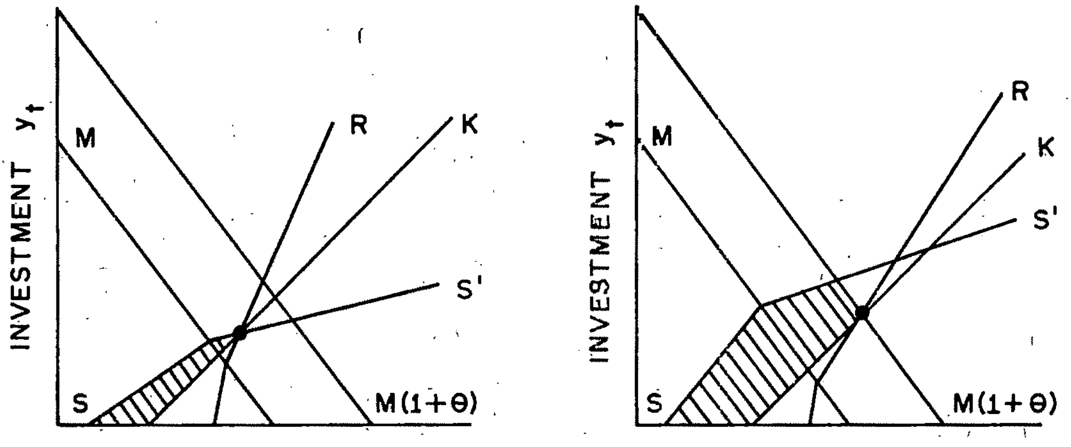
The decision problem for the firm is now represented as the maximization of (5) subject to (1)–(3) and (13)–(14). Several possible outcomes of this modified problem are shown in Figure 4. The feasible region is greatly compressed by the safety constraint which appears as the kinked line SS'. To the left of the kink, the constraint has slope

$$[(p_t' - (1 + i)c)(1 - \beta) + \beta q(\delta + i)]/q(1 + i)$$



PRODUCE AND INVEST
(a)
II'

PRODUCE, INVEST AND SAVE
(b)
VI



PRODUCE, INVEST AND BORROW
UP TO EXPECTED
CASH CONSTANT
(c)
VII

PRODUCE, INVEST AND BORROW
UP TO EXPECTED
BORROWING LIMIT
(d)
I'

FIGURE 4. DECISION RULES FOR THE FIRM WITH RISK AVERSION

which is the range in which borrowing does not occur. When borrowing takes place to the right of the kink, the slope is

$$[(\bar{p}_t' - (1+i)c)(1-\beta) + \beta q(\delta + i) - (r-i)c]/q(1+r)$$

P14418

The additional rules of thumb to which the various possible risk programming solutions correspond can be expressed as follows:

I′. *Produce, Invest, and Borrow Up To the Borrowing Constraint* if:

A. $p_t' > c(1 + r) + q(\delta + i)$

B. $\theta m < \{B(q + c) + (A + c(1 + i))k_{t-1}(1 - \delta)q + (A - q(1 + i))m_t\} \div \{A - q(1 + r) + c(i - r)\}$

for $m_t > (1 - \delta)ck_{t-1}$ or

$$\frac{(1 + \theta)m_t - (1 - \delta)k_{t-1}c}{q + c} < \frac{B + (A + c(i - r))k_{t-1}(1 - \delta) - m_t(i - r)}{A + c(i - r) - q(1 + r)}$$

for $m_t < (1 - \delta)ck_{t-1}$

In these and the following expressions

$$A = (p_t' - (1 + i)c)(1 - \beta) + \beta q(\delta + i)$$

and

$$B = \tilde{M} - m_t + iq(1 - \delta)k_t$$

Condition A asserts that investment is profitable, and condition B asserts that the borrowing limit is reached before the firm has invested up to the limit imposed by the safety margin. The two resource side conditions stem from the same difference in initial conditions on working and fixed capital alluded to in case VII below. The solution for $m_t > k_{t-1}(1 - \delta)c$ is shown in Figure 3d.

VI. *Produce, Invest, and Save* if:

A. $p_t' > c(1 + i) + (\delta + i)q$

B. $\dfrac{B + A(1 - \delta)k_{t-1}}{A - q(1 + i)} < \dfrac{m_t - c(1 - \delta)k_{t-1}}{q + c}$

C. $m_t > c(1 - \delta)k_{t-1}$

Condition A asserts that investment is profitable. Condition B means that the safety margin constraint is reached before the firm has invested all of its cash resources. Condition C asserts that the capital constraint is binding. This solution is illustrated in Figure 4b.

VII. *Produce, Invest, and Borrow Up To the Safety Margin Constraint* if:

A. $p_t' > c(1 + r) + q(\delta + r)$

B. $\{B(q + c) + (A + c(1 + i))k_{t-1}(1 - \delta)q + (A - q(1 + i))m_t\} \div \{A - q(1 + r) + c(i - r)\} < \theta m_t$

for $m_t > k_{t-1}(1 - \delta)c$ or

$$\frac{B + (A + c(i - r))(1 - \delta)k_{t-1} - m_t(i - r)}{A - q(1 + r) + c(i - r)} < \frac{(1 + \theta)m_t - c(1 - \delta)k_{t-1}}{q + c}$$

for $m_t < (1 - \delta)ck_{t-1} < (1 + \theta)m_t$

Condition A asserts that the internal rate of return on investment is greater than the borrowing rate of interest. Condition B means that the cash safety margin is exhausted before the firm has invested up to the limit of its borrowing capacity. There are two resource side conditions here reflecting the difference between the solution when the capital constraint is less or greater than the working capital constraint. The solution for the former case is illustrated in Figure 4c.

In the absence of risk averting behavior, the firm always invests, when it invests at all, up to the limit of its cash resources or borrowing capacity. When uncertainty of the future receipts is accounted for, however, guarding against possible future illiquidity becomes an important part of the firm's dynamic investment strategy. Investment in fixed capital requires an expenditure of current working capital which is in excess of the added receipts which that investment is expected to produce in the initial period. Investment requires a net cash drain in the amount $q(1+r)-c(i-r)-A$. It therefore may be rational to forego capital accumulation, even when investment is profitable, if the cash drain increases the probability of having a future shortage of working capital to an unacceptable level. Risk aversion is consistent with a firm simultaneously saving and investing, or borrowing less than the full amount available to it even when investment is profitable.

A change in the firm's estimate of uncertainty, σ, or in its minimum acceptable probability of having too little working capital, will definitely affect its investment behavior. An increase in uncertainty, for example, will increase the intercept of curve S, thus shifting it to the right in Figure 4. This will reduce investment in all cases where the safety constraint is binding.

Risk aversion could be thought of as an alternative to, or an example of, a cost of adjustment. It is a factor explaining the speed of adjustment of an industry or a firm when it is not in long-run equilibrium. Interestingly enough, on the industry level this cautious behavior reduces the probability of unstable cycles around equilibrium, because it reduces the size of the periodic change in output and capacity which each firm in the industry will undertake.

IV. A Comparison with Alternative Theories

How does our myopic optimizing model compare with other micro-economic theories of investment? A number of recent investigators have found that the firm's financial position has an important influence on its investment behavior (W. H. Locke Anderson, John Lintner, and John Meyer and Edwin Kuh). The model presented in this paper provides a rationale for this result.

The neoclassical theory of investment as developed by Dale Jorgenson and Jorgenson and James Stephenson represents an alternative explanation for investment by the firm. In that theory the firm forecasts the entire future, determines an optimal capital stock and then invests as fast as is technologically feasible to achieve this optimum. The speed of adjustment is not variable and investment at less than the rate imposed by a borrowing ceiling never appears rational. In order to derive an optimal capital stock, it is assumed that the current production level is the long-run optimum. (See Jorgenson, p. 249.) Even under assumed stationary market conditions this assumption is inconsistent since as the capital is installed, the output level will change. Essentially therefore their investment plan is based on a possibly suboptimal rule for determining the succession of capital stocks. By way of contrast the business firm of this paper makes only a shortsighted forecast. It computes a first step on some trajectory which, as we have seen, may or may not converge to equilibrium and long-run purely competitive optimality. If it converges, it does so according to quite simple rules, akin to trial and error, that do not estimate equilibrium at all. Furthermore our analysis incorporates risk, uncertainty, and liquidity, three factors which are assumed away in the standard treatment.

Another segment in the literature (John Gould, Robert Lucas, and Arthur Treadway), adds a dimension to the neoclassical theory by allowing the firm's computation of an investment plan to depend on the cost of the firm's adjustment to a larger scale of operations. This theory involves the use of intertemporally optimal strategies based on perfect foresight. But, if allowance is made for the cost of information and risk, the optimal path might be closely approximated by the myopic rules discussed here. There no doubt exist conditions of demand and cost where an industry of myopic optimizers would converge to the long-run optimal capital stock along an optimal adjustment path.

Whether or not a truly optimal result is brought about in reality is an important empirical question. But the actual path, whatever its properties, seems to us best described as the result of a cautious groping toward an unknown equilibrium, in which future developments as they become known through market feedback tell the firm whether its goal has been passed in the fog or still lies somewhere ahead. Levels of investment and the speed of adjustment may be better explained by myopic optimizing with imperfect information and allowance for risk, than by the presently popular neoclassical model.

REFERENCES

W. H. L. Anderson, "Business Fixed Investment: A Marriage of Fact and Fancy," in R. Ferber, ed., *Determinants of Investment Behavior,* New York 1967.

W. J. Baumol and R. A. Quandt, "Investment and Discount Rates Under Capital Rationing," *Econ. J.*, June 1965, *75*, 317–29.

——— and ———, "Rules of Thumb and Optimally Imperfect Decisions," *Amer. Econ. Rev.*, Mar. 1964, *54*, 23–46.

S. Carlson, *A Study on the Pure Theory of Production,* London 1939.

R. H. Day, "Profits, Learning and the Convergence of Satisficing to Marginalism," *Quart. J. Econ.*, May 1967, *31*, 302–12.

———, **D. J. Aigner, and K. R. Smith**, "Safety Margins and Profit Maximization in the Theory of the Firm," *J. Polit. Econ.*, Dec. 1971, *79*, 1293–1302.

——— **and R. Benson**, "A Short-Sighted Growth Model for the Firm and Industry," Univ. Wis., S.S.R.I., Workshop pap. no. 6909, July 1969.

——— **and E. H. Tinney**, "A Generalized Cobweb Theory of Economic Growth, Decay and Cycles," *Econ. J.*, Mar. 1969, *79*, 90–108.

J. P. Gould, "Adjustment Costs in the Theory of Investment of the Firm," *Rev. Econ. Stud.*, Jan. 1968, *35*, 47–57.

D. W. Jorgenson, "Capital Theory and Investment Behavior," *Amer. Econ. Rev. Proc.*, May 1963, *53*, 247–60.

——— **and J. A. Stephenson**, "Anticipations and Investment Behavior in U.S. Manufacturing, 1949–1960," *Econometrica*, Apr. 1967, *35*, 169–221.

J. Lintner, "Corporation Finance: Risk and Investment," in R. Ferber, ed., *Determinants of Investment Behavior,* New York 1967.

R. E. Lucas, "Optimal Investment Policy and the Flexible Accelerator," *Int. Econ. Rev.*, Feb. 1967, *8*, 78–85.

J. Meyer and E. Kuh, *The Investment Decision: An Empirical Study*, Cambridge 1959.

A. D. Roy, "Safety First and the Holding of Assets," *Econometrica,* July 1952, *20*, 431–38.

A. B. Treadway, "On Rational Entrepreneurial Behavior and the Demand for Investment," *Rev. Econ. Stud.*, Apr. 1969, *36*, 227–239.

J. B. Williams, "The Path to Equilibrium," *Quart. J. Econ.*, May 1967, *81*, 241–56.

International Capital Movements and the Patterns of Economic Growth

By YUSUKE ONITSUKA*

The purpose of this paper is to extend recent contributions in the theory of international capital movements and develop a more general approach by which the long-run growth patterns of an open economy, including those of a "mature creditor country,"[1] can be analyzed. Since Roy Harrod's emphasis on a dynamic approach to the theory of foreign investment, Evsey Domar, Harry Johnson, George Borts (1964, 1971), James Hanson and Philip Neher, Akihiro Amano, and Koichi Hamada have developed and elaborated a dynamic theory. Most of these contributions, however, confined the discussion to the analysis of the final stationary state or golden rule path of an open economy.[2] The effect of international capital movements upon an open economy in the various stages of economic growth have not, therefore, been adequately examined.[3] In particular the problem of a mature creditor country remains largely unexplored.[4]

The clues to a more general approach lie in the recent contributions by Lloyd Metzler and Hirofumi Uzawa (1968). In his lucid analysis of the short-run adjustment mechanism of an open economy, Metzler explicitly identifies the Keynesian marginal efficiency of investment as the supply of new securities, the saving as demand for them and excess saving as excess demand for them (capital outflows). Uzawa pointed out that the approach based on the schedule of marginal product of capital (marginal output-capital ratio) alone[5] could be misleading in the case of portfolio investment and suggested an analytical framework in which the schedule of savings and marginal efficiency of investment (as distinguished from marginal product of capital) play a central role.[6] The recent developments in the dynamic theory of investment made by Robert Lucas, Jr., Arthur Treadway and Uzawa (1969) have also provided the missing link between the desired level of

---

* Economist, International Monetary Fund. This paper is a compendium of my unpublished doctoral dissertation. The study was supported by a Ford Doctoral Dissertation Fellowship. I am grateful for the guidance and encouragement given by H. Uzawa, R. A. Mundell, H. G. Johnson, and R. Parks. I would also like to thank Franklin Mills, and Hiroshi Ono whose comments and suggestions were very helpful. I also received helpful comments from the participants including Harl Ryder in the symposium on "Uncertainty and/or Capital" held at Yale University in 1968.

[1] A creditor country whose investment income and amortization funds become greater than the volume of capital exports and reinvestment. See also fn. 14.

[2] There are a few exceptions; see Hamada, and Hanson and Neher.

[3] One of the limitations of previous contributions is that they do not take into consideration the marginal efficiency of investment which is the prime factor governing domestic capital accumulation. As a result, the mechanism of the growth of an open economy whose level of capital accumulation is below or above that of equilibrium has not been adequately treated. For example, there is no adjusting mechanism in earlier works other than instantaneous adjustment of capital stock if the equilibrium growth is disturbed by sudden changes in parameters such as world rate of interest. This, however, involves conceptual difficulty because one of the basic economic realities which give rise to capital movements is the fact that the levels of capital stock cannot be adjusted instantaneously and that there is a persistent difference in the rates of return to real capital among regions and countries.

[4] There are only few exceptions; see, for example, Domar, also see fn. 13.

[5] This approach is represented by Murray Kemp, ch. 13–14, and G. D. A. MacDougall.

[6] For the conceptual difference between the marginal efficiency of investment and the marginal product of capital, see Abba Lerner. See also Onitsuka, pp. 12–23.

investment and the continuously growing capital-labor ratios.

In this paper, I utilize Uzawa's analytical framework and attempt to integrate some aspects of the abovementioned contributions. The long-run growth patterns of an open economy will be discussed in terms of various phases of economic growth characterized by levels of the capital-labor ratio and per capita indebtedness and of the direction of capital flows. The paper will discuss the relative size of capital flows and interest receipts (payments) in each growth phase. In particular, it will be shown that a mature creditor country whose interest receipts exceed capital exports is in the final subphase to which an economy converges when the savings ratio is larger than a certain value. The discussion is confined to international capital movements in the context of portfolio investment as in the abovementioned contributions. Section I presents the theoretical framework including the model. The factors determining the direction and magnitude of international capital movements will also be discussed. In Section II, I analyze the growth patterns of an open economy. The long-run growth patterns depend upon the magnitude of the saving ratio relative to other parameters and are classified into three cases according to the nature of the steady stationary state to which the growth paths converge. These are: 1) high saving ratio case; 2) low saving ratio case I; and 3) low saving ratio case II. For each case, various growth paths of an open economy will be examined in terms of the abovementioned characterization.

I. Theoretical Framework

A. *The Model*

As mentioned in the introduction, the most important characteristics of international capital movements in the context of portfolio investment is the international exchange of buying power (claims on currently produced commodities) for securities (claims on future income streams).[7] It is, therefore, important to treat this aspect explicitly in the analysis of capital movements and to make a conceptual distinction between international capital movements and changes in capital stock (real capital) which may or may not be related to capital movements[8] as such.

Let us consider a small country which is faced with a perfectly competitive international capital market. The country produces only one kind of commodity which can be either consumed or invested as capital goods. With a one-commodity model, we can simplify the analysis by concentrating on the effects of capital movements upon capital accumulation and other macroeconomic variables. For our purpose, the trade of commodities, including that of investment goods, need not be treated explicitly because of the primarily financial nature of capital movements.

Let B be the value of net holdings of foreign securities by a small country, $\bar{i}$ the prevailing world rate of interest, and T be the interest receipts. Then we have

$$T = \bar{i}B \tag{1}$$

And, if the world rate of interest remains constant over time[9]

---

[7] This financial transaction can be also referred to as: 1) foreign investment (excluding direct investment); 2) lending (borrowing); and 3) export (imports) of capital (excluding direct investment).

[8] Bertil Ohlin and Carl Iversen made the clear distinction. They identify capital movements as movements of "buying power" (Ohlin) or "waiting" (Iversen) and distinguished them from the shifts in real capital, calling the latter "factor movements." J. E. Meade and Mundell also made a similar distinction. MacDougall and Kemp, however, do not seem to make this distinction. Furthermore, the exports of capital goods (machines) or renting them out are not exports of capital because in neither case is sales of securities or borrowing involved. The former is merely exports of commodities and the latter is sales of service of machines. For further discussion of the concept of capital movements, see Onitsuka, pp. 12–23, and Ohlin, chs. 20 and 21; Iversen, pp. 19–20; and Meade.

[9] If the world rate of interest rate is changing, the

$$\dot{B} = T/\bar{i} \tag{2}$$

The increase over time of the value of foreign securities is equal to the excess of savings S over investment I, or excess demand for new securities. That is,

$$\dot{B} = S - I \tag{3}$$

where savings depend upon the level of income Y and the saving ratio s which is assumed to be a function of the rate of interest,

$$S = s(\bar{i})Y = s(\bar{i})(F(K, L) + T) \tag{4}$$

where $F(K, L)$ stands for national product as a function of capital and labor. Investment is determined by the following schedule of marginal efficiency of investment[10]

$$I = \phi(\bar{i}, r)K \tag{5}$$

where ϕ is the investment expenditure per unit of capital stock, and r the expected rate of profit.

Together with equations (2), (4), (5), equation (3) indicates how the rate of increase in interest receipts $\dot{T}$ is determined for the given K and T at each moment in time

$$\dot{T} = \bar{i}s(F(K, L) + T) - \bar{i}\phi K \tag{6}$$

The other two basic variables, capital stock and the number of workers are increasing according to the following equations:

$$\dot{K} = zK \tag{7}$$

$$\dot{L} = nL \tag{8}$$

The above differential equations, (6), (7), and (8), can be reduced to a two-equation system by using per capita variables.[11]

$$\dot{q} = (\bar{i}s - n)q + \bar{i}(sf(k) - \phi(\bar{i}, r)k) \tag{9}$$

$$\dot{k} = (z - n)k \tag{10}$$

where q is per capita interest receipts, T/L, and k the capital-labor ratio, K/L.

B. *The Factors Governing International Capital Movement*

In describing the factors governing the capital flows, it is convenient to make a conceptual distinction between two aspects. Savings, investment, and capital flows will be called "the flow aspect" because they have a flow dimension, i.e., change per unit of time. The levels of capital accumulation and external indebtedness will be called "the stock aspect" since they have a stock dimension. As equation (6) shows, it is the flow aspect, not the stock aspect, which directly determines the direction and magnitude of capital flows. The stock aspect influences capital flows only to the extent to which it affects the flow aspect. The manner in which the stock aspect influences the flow aspect depends upon the world rate of interest, time preference and the marginal efficiency of investment which is affected by the endowment of entrepreneurial capacity to increase capital stock. This point needs to be emphasized because it has not received adequate attention. Its implication is that one cannot analyze international capital movements in the context of portfolio investment without explicitly taking into account the flow aspect. For example, it is possible that a "capital-poor" country, i.e., a country with low

---

transformation does not hold. That is to say, $dB/dt = (1/i)(dT/dt) - (di/dt)(T/i^2)$. The second term of the right-hand side stands for capital gain or loss due to the changes in interest rate.

[10] For given world rate of interest and rate of profit, each firm chooses the optimal rate of increase in capital stock, z, maximizing the present discounted value of net profit. In turn the level of z determines the level of investment expenditure ϕ; i.e., $\phi = \phi(z) = \phi(i, r)$. Note also that $\partial\phi/\partial i < 0$, $\partial\phi/\partial r > 0$. For the details of derivation of the investment schedule, see Uzawa (1969).

[11] Equation (6) is transformed to equation (9) utilizing the following relation: $\dot{T}/L = \dot{q} + qn$. Similar equations have been derived by a number of authors of the abovementioned contributions. However, these authors did not incorporate explicitly the schedule of marginal efficiency of investment as mentioned earlier.

capital-labor ratio, exports capital to a "capital-rich" country if the effects of a high saving ratio and a small endowment of entrepreneurial capacity in the capital-poor country dominates that of the high marginal product of capital in the security market. In other words, a country which is poor in capital stock can be a capital-rich country in terms of portfolio capital in the sense that it has an excess demand for new securities at the world equilibrium rate of interest.[12]

The factors which dictate the capital flows can be analyzed by differentiating equation (6) with respect to each of these factors.

$$(11)\quad \frac{\partial B}{\partial S} = 1 > 0$$

$$(12)\quad \frac{\partial \dot{B}}{\partial I} = -1 < 0$$

$$(13)\quad \frac{\partial \dot{B}}{\partial K} = s\frac{\partial F(K, L)}{\partial K} - \frac{\partial \phi}{\partial r}\frac{\partial^2 F}{\partial K^2}K - \phi \gtrless 0$$

$$(14)\quad \frac{\partial \dot{B}}{\partial T} = s > 0$$

$$(15)\quad \frac{\partial \dot{B}}{\partial i} = \frac{\partial s}{\partial i}[F(K, L) + T] - \frac{\partial \phi}{\partial i}K > 0$$

$$(16)\quad \frac{\partial \dot{B}}{\partial s} = F(K, L) + T > 0$$

The inequalities (11) and (12) indicate that an increase in savings (investment), *ceteris paribus*, increases (decreases) capital outflows by an equal amount. The inequality (13) implies that the effect of capital accumulation on the capital outflow cannot be determined a priori. An increase in capital stock increases savings through higher income but the effect upon investment is not certain, i.e., while a lower rate of profit will reduce investment per unit of capital the scale factor K is larger. However, the primary effect of capital accumulation upon investment is likely to be a reduction caused by a fall in the rate of profit. It is assumed in later discussion that an increase in capital stock, other factors kept constant, results in an increase of capital outflows. The inequality (14) shows that an increase in the value of foreign securities or interest receipts leads to an increase in capital outflows through an increase in aggregate savings. The inequality (15) indicates that an increase in the world rate of interest increases the country's capital exports. The inequality (16) indicates that an increase in the saving propensity or a reduction in time preference always increases capital outflows. Finally, the effect of an increase in the supply of entrepreneurial capacity, which is not shown among the above inequalities, is uncertain because while it decreases the investment expenditure required to achieve a given rate of increase in capital stock, the firms' desired growth rate of capital stock will also increase due to the greater profitability of investment.

C. *Characterization of the Growth Phases*

The various phases of economic growth of an open economy can be characterized in terms of the stock aspect and the flow aspect. The first characteristic is the level of foreign securities held by the residents of the economy. It indicates the net position of an economy's accumulated capital transactions with the rest of the world. An economy is called a "creditor" if it holds a positive level of foreign securities, i.e., $T/i = B > 0$, and a "debtor" if the rest of the world is holding securities issued by the home country's firms, i.e., $T/i < 0$.

The second characteristic is the level of capital flows or excess demand for new securities. An economy will be called a

---

[12] For more detailed discussion, see Onitsuka, pp. 12–23.

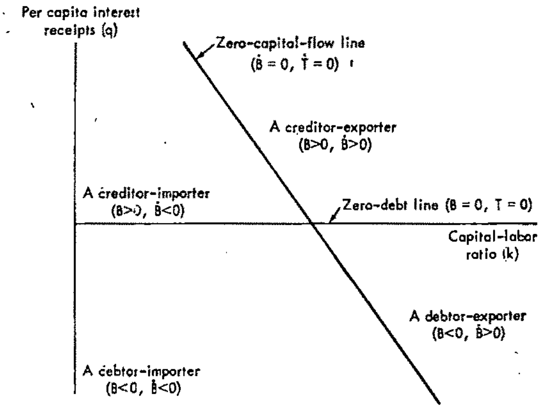

Per capita interest receipts (q)
Zero-capital-flow line ($\dot{B} = 0$, $\dot{T} = 0$)
A creditor-exporter ($B>0$, $\dot{B}>0$)
A creditor-importer ($B>0$, $\dot{B}<0$)
Zero-debt line ($B = 0$, $T = 0$)
Capital-labor ratio (k)
A debtor-exporter ($B<0$, $\dot{B}>0$)
A debtor-importer ($B<0$, $\dot{B}<0$)

FIGURE 1

"capital-exporter" if it has an excess demand for new securities, i.e., $\dot{T}/i = \dot{B} > 0$, or $S > I$, at the world rate of interest, and a "capital-importer" if it has excess supply of securities $\dot{T}/i = \dot{B} < 0$, or $S < I$.

The characterization of the growth phases is presented in Figure 1. The vertical axis in the figure indicates the level of per capita interest receipts and the horizontal axis indicates the level of the capital-labor ratio k. The k-axis is also called the zero-debt line because the holding of foreign securities on this line is zero, i.e., $B = T/i = 0$. The zero-capital flow line along which the level of capital flow is zero, i.e., $\dot{B} = 0$ has a negative slope because an increase in capital-labor ratio will increase capital outflow due to higher income and a lower rate of profit. Therefore, a decline in per capita interest receipts is required to maintain the capital flow at its original level.[13] In the area above the zero-capital flow line the holdings of foreign securities are increasing, i.e., $\dot{B} > 0$ and they are decreasing below it, i.e., $\dot{B} < 0$. By combining those two characteristics, we have four types of growth phase. An economy will be called a "creditor-importer" if it is a creditor in terms of the stock aspect, i.e., $B > 0$, but a capital-importer in terms of the flow aspect, i.e., $\dot{B} < 0$. Likewise we have a "debtor-importer," a "debtor-exporter" and a "creditor-exporter" as shown in Figure 1.

Finally, it is possible further to divide each growth phase into two subphases by examining the relationship between the volume of capital exports (imports) and that of inflow of funds due to investment income. This extension turns out to be useful in analyzing the abovementioned problem of a mature creditor or the "import balance" problem.[14]

The lines which divide each phase into two subphases are obtained by the following relationship in which the volume of capital flows equals that of interest payments (receipts).

$$(17) \qquad |R_n| = \left|\frac{iB_f}{\dot{B}_f}\right| = \left|\frac{T}{\dot{T}/i}\right| = 1$$

where R_n is the ratio of the inflow (outflow) of funds due to interest payments (receipts) to flows of funds due to capital exports (imports). This ratio will be called the inflow-outflow ratio.[15] In the case of a creditor-exporter and debtor-importer, the ratio is positive because investment income and portfolio capital flows in an opposite direction and therefore have the same signs in our definition of variables. Hence equation (17) in this case becomes

---

[13] See the Appendix, Section 1.

[14] The existence of this phenomenon was known as early as Cairnes and Marshall. In the 1940's, it received attention again when Domar, Buchanan, Hinshaw, and others addressed themselves to the likely consequences of continued foreign aid and investment. However, the emphasis was upon the impacts of import balance upon domestic employment and most of the analysis is in terms of numerical examples. In particular, Domar attempted to develop a more general approach by examining the relative magnitudes of the factors which determine the growth path of a creditor country. These contributions, however, isolate the foreign sector from domestic capital accumulation and do not treat this aspect as a particular stage of economic growth of an open economy. See Domar; Buchanan, pp. 166–80, 206–39; Hinshaw, pp. 661–71; and Hal Lary.

[15] In Domar's case, the inflows of funds include not only interest receipts but also amortization funds. The outflow of funds, therefore, becomes gross foreign investment which includes reinvestment of amortization funds. The inflow-outflow ratio (R) becomes $(iB+A)/(\dot{B}+C)$ where A and C stand for amortization and reinvestment, respectively. By dividing both the numerator and denominator by B, we obtain $R = (i+a)/(b+c)$ where a, b, and c are amortization ratios (A/B), the growth rate of foreign securities $(\dot{B}/B)$ and the ratio of reinvestment to the value of foreign securities. This

$$R_n = iT/\dot{T} = 1 \tag{18}$$

Taking into account equation (6),

$$q = \frac{1}{1-s}\,[sf(k) - \phi k] \tag{19}$$

In the case of a creditor-importer and debtor-exporter, the ratio is negative because both types of funds flow in the same direction. Hence equation (17) becomes

$$R_n = -1 \tag{20}$$

or

$$q = \frac{-1}{1+s}\,[sf(k) - \phi k] \tag{21}$$

The lines indicating equality of capital flows and interest receipts (payments) and the subphases generated by these lines are shown in Figure 2.

In the case of a creditor-importer (debtor-exporter), aggregate interest receipts (payments) which are initially larger than capital inflows (outflows) eventually become less than capital inflows (outflows), since interest receipts (payments) are declining in this phase. A country in the early stage of the phase is called an "immature creditor-importer" (an "immature debtor-exporter"). Likewise we can have an "immature debtor-importer" (an "immature creditor-exporter") if interest payments (receipts) are less than capital inflows (outflows) and a "mature debtor-importer" (a "mature creditor-exporter") if interest payments (receipts) are greater than capital inflows (outflows). Those characteristics of the growth phases and subphases are summarized in Table 1.[16] Of particular importance is the subphase representing the mature creditor-exporter which is marked by a hatched area. It is the phase in which interest receipts exceed the volume of capital exports and the economy enjoys a net inflow of funds. Symmetrically, a mature debtor-importer suffers from net outflow of funds in spite of capital imports.

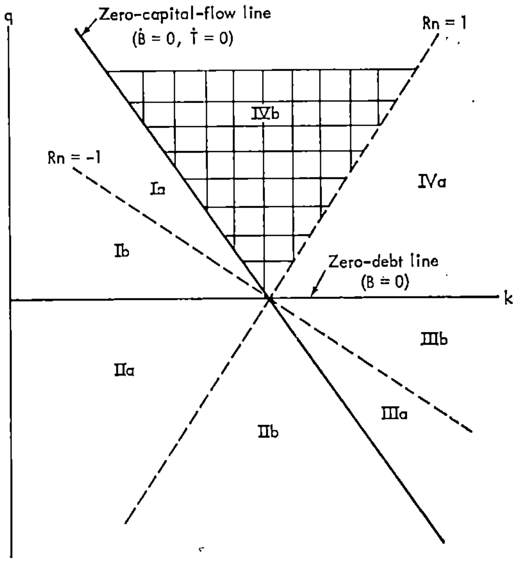


FIGURE 2

II. Patterns of Economic Growth

A. *The High-Saving Ratio Case*

This is the case where the country has a saving ratio higher than the ratio of the growth rate of population and the world rate of interest, i.e., $s > n/i$. Various possible growth paths are shown in the phase diagram (Figure 3)[17] which is obtained by superimposing the $\dot{k} = 0$ line and the $\dot{q} = 0$ line onto Figure 2.

Path *A* in the phase diagram shows the

---

gross inflow-outflow ratio, R, however, is analytically equivalent to our net inflow-outflow ratio, R_n, in the small country case because the amortization it receives has to equal its reinvestment to the rest of the world.

[16] For a similar but somewhat different classification of development stages, see C. P. Kindleberger.

[17] For the explanation of Figure 3, see the Appendix, Section 2.

Table 1—The Summary of Phases of Economic Growth

	I		II		III		IV	
Phases	Creditor-Importer		Debtor-Importer		Debtor-Exporter		Creditor-Exporter	
Subphases	Ia	Ib	IIa	IIb	IIIa	IIIb	IVa	IVb
	Immature creditor-importer	Mature creditor-importer	Immature debtor-importer	Mature debtor-importer	Immature debtor-exporter	Mature debtor-exporter	Immature creditor-exporter	Mature creditor-exporter
	Interest receipts are greater than capital inflows	Interest receipts are less than capital inflows	Interest payments are less than capital inflows	Interest payments are greater than capital inflows	Interest payments are greater than capital outflows	Interest payments are less than capital outflows	Interest receipts are less than capital outflows	Interest receipts are greater than capital outflows

typical path of an open economy with a high saving ratio. In subphase IIa, a capital-poor country opening the economy to the world capital market first becomes an immature debtor-importer by importing capital. The volume of capital imports exceeds that of interest payments because a low capital-labor ratio gives rise to a high marginal product of capital, thereby causing a large volume of capital imports relative to foreign debt which is only beginning to be accumulated.

As the domestic capital accumulation is accelerated by capital imports, the marginal product of the capital stock declines. Capital imports decrease relative to interest receipts owing to a higher rate of domestic savings and to declining investment. The volume of capital imports becomes first equal to and then less than that of the interest payments as the economy hits the $R_n = 1$ line and becomes a mature debtor-importer (subphase IIb).

By the time the economy arrives at the zero-capital flow line, capital imports are reduced to zero. In the subphase IIIa, the direction of capital flows reverses. The economy now becomes an immature debtor-exporter with capital exports still less than interest payments to foreigners.

In subphase IIIb, the volume of capital exports becomes first equal to (the $R_n = -1$ line) and then greater than that of debt servicing to the rest of the world. As the cumulative effect of capital exports manifests itself, the debtor's position is reduced to zero (zero-debt line) and the economy turns into a creditor becoming an immature creditor-exporter (subphase IVa). Its interest receipts are still less than that of the capital outflows.

Finally, the economy hits the $R_n = 1$ line again and moves into subphase IVb of a mature creditor-exporter where interest receipts exceed the volume of capital exports. The economy will remain as a mature creditor and enjoy inflow of funds forever in spite of capital outflows. Path *B* indicates that an immature creditor-importer characterized by a high level of foreign security holdings and a low capital-labor ratio can become a mature creditor-exporter without being a debtor-importer or a debtor-exporter.

Path *C* shows that a mature creditor-exporter remains in the same subphase forever, although initially it could suffer from a decline in per capita interest receipts.

Two characteristics of the growth pattern resulting from a high saving ratio should be mentioned. First, no matter where they start, the paths end up with a mature creditor-exporter and the per capita income increases indefinitely due to accumulation of foreign securities. Secondly, the per capita income $f(k)+q$ is always increasing along the paths in the positive income region. A path like D which would lead the economy to a negative-income region due to increasing indebtedness does not exist,[18] unless the initial condition lies in the negative-income region.

Most of the previous contributions confine themselves to the steady stationary state as mentioned earlier. However, as can be easily seen from Figure 3 analysis confined to the growth paths in the stationary state permit us to examine only growth paths in the subphases IIIa, IIIb, IVa, and IVb, leaving out those in the phases of a creditor-importer (Ia and Ib) and a debtor-importer (IIa and IIb). In particular, the growth process of an open economy in phases II and III is worth treating explicitly since it corresponds not only to the early development stages of some of the developed countries but also to the current status of developing countries.

As the economy approaches the stationary state, the inflow-outflow ratio converges to the reciprocal of the saving ratio.

$$\lim_{t\to\infty} R_n(t) = \frac{1}{s} \tag{22}$$

It also follows that the rate of growth of aggregate interest receipts and that of the value of foreign securities converge to the product of the saving ratio and the world rate of interest.

$$\lim_{t\to\infty} \frac{\dot{T}(t)}{T(t)} = \lim_{t\to\infty} \frac{\dot{B}(t)}{B(t)} = is \tag{23}$$

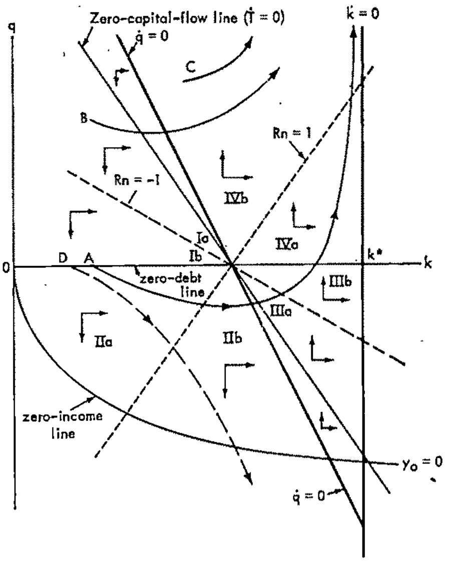


FIGURE 3

B. *The Low-Saving Ratio Case I*

This is a case in which the saving ratio is small but larger than the investment-output ratio at the stationary state, i,.e. $n/i > s > \phi(n)k^*/f(k^*)$.[19] It is an interesting case because this range of saving ratios still makes the economy a creditor-exporter at the later stage of economic growth in spite of the fact that the saving ratio is less than n/i. Earlier contributions have not treated this case presumably because they do not make a conceptual distinction between the rate of interest (cost of borrowing) and the marginal product of capital (rate of profit).

The growth paths of the low-saving ratio case are depicted in the phase diagram

---

[18] This is an interesting finding since it shows that the unstable nature of the system due to the high saving ratio is asymmetric. For the details of the proof that path D does not exist, see the Appendix Section 3, and Onitsuka, pp. 95–96.

[19] The investment-output ratio at the stationary state is always smaller than n/i. The inequality $s > \phi(n)k^*/f(k^*)$ is related to the condition under which the economy will still be a creditor at the stationary state k^*. The investment-output ratio could be approximated by n/r^* where r^* is the marginal product of capital of the stationary state. It cannot be determined a priori, however, that the investment-output ratio is greater or less than n/r^*. It may be noted also that $n/i > n/r^*$ because $r^* > i$.

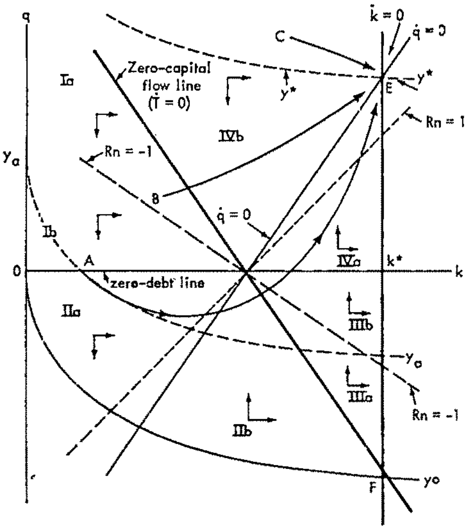


FIGURE 4

(Figure 4).[20] A capital-rich economy which does not have sufficiently high level of foreign securities follows a path similar to those of the high-saving ratio case except in the final subphase of a mature creditor-exporter (paths A and B). The major difference from the previous case is that per capita interest receipts and hence income do not increase indefinitely but converge to certain levels. The levels of those variables are the following:

$$(24)\quad q^* = -\frac{i}{(is-n)}\,[sf(k^*) - \phi(n)k^*]$$

$$(25)\quad y^* = f(k^*) + q^*$$

$$= -\frac{i}{(is-n)}\,nk\left[\frac{f(k^*)}{ik^*} - \frac{\phi(n)}{n}\right]$$

As can be seen in the figure, whether or not the stationary state E belongs to the subphase of a mature creditor-importer depends crucially upon the steeepness of the slope of the $\dot{q} = 0$ line relative to that of the $R_n = 1$ line. It can be safely concluded, however, that E lies in the subphase of a mature-creditor importer because the slope of the former is less steep than that of the latter so long as the world rate of interest is greater than the rate of population growth.

Path C indicates that a creditor-exporter with high foreign security holdings suffers from a continuous decline of per capita income until it reaches the stationary state. This occurs because the saving ratio which is the prime factor determining the rate of increase in foreign security holdings is not large enough to offset the impacts of population growth upon per capita security holdings. Thus, any economy which happens to be in the region above the iso-income curve y^* eventually suffers an absolute decline in per capita income. However, a path like D which would lead to negative income due to indebtedness does not exist because of the stable nature of the system.

Finally, the above level of per capita income in the stationary state is different from that of a closed economy[21] but similar to those of an open economy obtained by other authors.[22] First of all, savings do not have to be equal to investment due to the presence of international capital movements. Secondly, the saving ratio is determined independently by the world rate of interest in the case of a small country facing the world capital market. The stationary income y^*, however, has the same properties as that of a closed economy in terms of the effects on it of changes in saving ratio and the rate of growth of population. The stationary-state income is increased by an increase in the propensity to save and decreased by an increase in the rate of growth of population.

$$(26)\qquad \frac{\partial y^*}{\partial s} > 0$$

---

[20] For the explanation of Figure 4, see the Appendix, Section 4.

[21] See Solow and Swan.

[22] See Borts, Amano, and Neher.

$$\frac{\partial y^*}{\partial n} < 0\,. \tag{27}$$

Also, an increase in the supply of entrepreneurial capacity will raise the stationary-state income.[23] The effect of changes in the world rate of interest cannot be determined a priori.[24]

The value of the inflow-outflow ratio converges to the ratio of the world interest rate to the growth rate of population, since the aggregate interest receipts are growing at the same rate as the population.

$$\lim_{t\to\infty} R_n(t) = \lim_{t\to\infty} \frac{T}{\dfrac{\dot{T}}{i}} = \frac{i}{n} \tag{28}$$

C. *The Low-Saving Ratio Case II*

This is a case where the economy's saving ratio is smaller than the investment-output ratio at the stationary state, i.e., $s<\phi(n)k^*/f(k^*)$.[25] The growth paths of the economy are shown in the phase diagram (Figure 5).[26] All the growth paths now converge toward a debtor position as indicated by the negative level of interest receipts *E* in the figure. Path *A* for example shows that as in previous cases a capital-poor economy first becomes an immature debtor-importer (IIa) and then a mature debtor-importer (IIb) as indebtedness increases. The economy, however, will never become a capital-exporter or a creditor at a later stage because the stationary state lies below the zero-capital-flow line. The value of per capita income and that of the inflow-outflow ratio at the stationary state are the same as in the low-saving ratio case I. The level of indebtedness, however, is never large enough to make that level of per capita income negative even in this low-saving ratio case.[27]

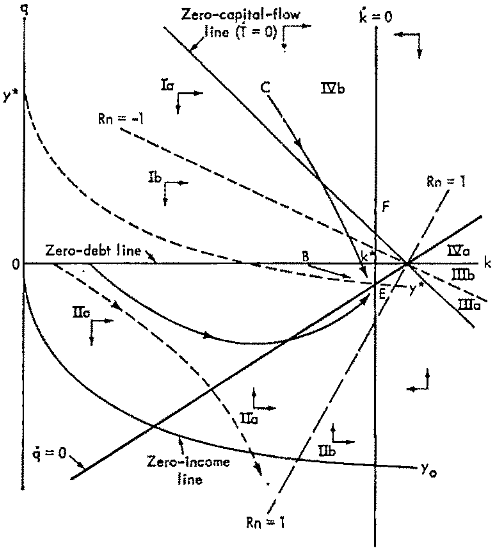


FIGURE 5

III. Summary and Conclusion

The paper attempts to develop a general theory of international capital movements in the context of portfolio investment by integrating recent developments in the dynamic theory of investment. The paper analyzes the growth patterns of a small open economy in terms of various phases of economic growth characterized by the levels of capital flows, indebtedness and domestic capital accumulation. The main thrust of this attempt is to make a conceptual distinction between marginal efficiency of investment and marginal product of real capital, identifying the former as supply of new securities.

The long-run patterns of economic growth are classified into three cases according to the magnitude of the saving ratio relative to other parameters. In the high-saving ratio case, the per capita income increases indefinitely due to accumulation of foreign securities as the capital-labor ratio converges to the stationary state. In the other two cases, it converges to a certain positive value and it approaches a creditor position in the low-saving ratio case I and to a debtor position in the low-saving ratio case II. For each case, the typical paths are examined in terms of the direction of capital flows and holdings of foreign securities.

---

[23] $\partial y^*/\partial\alpha = \partial y^*/\partial\phi(n,\alpha)\cdot\partial\phi(n,\alpha)/\partial\alpha > 0$, where α is a shift factor of the Penrose curve. An increase in the supply of entrepreneurial capacity is interpreted as a reduction of investment expenditure required to achieve the same rate of growth of capital stock.

[24] For further discussion of this effect, see Onitsuka, pp. 51–52.

[25] For the implication of this condition, see fn. 19.

[26] For the explanation of Figure 5, see the Appendix, Section 5.

[27] For the proof, see the Appendix, Section 6.

In all three cases, an open economy with a low-capital-labor ratio typically imports capital thus becoming a debtor at the early state of economic growth. In the first two cases it then starts exporting capital after a certain level of capital accumulation while maintaining a debtor position. It eventually becomes a creditor at the later stage as it continues to export capital. In the third case, however, it never becomes a capital-exporter or a creditor.

Each of the abovementioned growth phases is further divided into two subphases according to the magnitude of the volume of capital exports (imports) relative to interest payments (receipts). The paper has shown that an open economy always converges to the final subphase of a mature creditor if the saving ratio is greater than the investment-output ratio at the stationary state, i.e., even if the saving ratio is smaller than the ratio of the growth rate of population to the world rate of interest.

APPENDIX

1). From equation (6) in the text, the zero-capital flow line is expressed by $\dot{T}/i=0$ or $0=s(i)(F(K, L)+T)+\phi(i, r)K$, which can be expressed in terms of per capita variables, $q=[\phi(i, r)k-sf(k)]1/s$. The slope of this curve is given by $dq/dk=-1/s(sf'+\phi/rf''k-\phi)<0$.

2). The $\dot{k}=0$ line on which the capital-labor ratio remains constant is vertical in the (k, q) plane for the condition $\dot{k}=0$ is independent of per capita interest receipts. (From equation (10), $\dot{k}=0$ implies $z(i, r)=n$, where $r=f'(k)$.) The capital-labor ratio is increasing at the left-hand side of the $\dot{k}=0$ line because lower capital-labor ratios make the marginal product of capital high enough to make the desired rate of capital accumulation greater than the rate of population growth. This will be shown by differentiating equation (10) at $\dot{k}=0$, namely,

$$\frac{\partial \dot{k}}{\partial k}=\frac{\partial z}{\partial r}f''k<0$$

The $\dot{q}=0$ line on which per capita interest receipts remain constant is negatively sloped but steeper than the zero-capital flow line. The slope of the $\dot{q}=0$ line is obtained by differentiating equation (9) with respect to k at $\dot{q}=0$. That is,

$$\frac{dq}{dk}=-\frac{i}{is-n}\left(sf''-\frac{\partial\phi}{\partial r}f''k-\phi\right)<0$$

The absolute value of the above slope is greater than that obtained above in Section 1). The system is unstable with respect to q because $\partial\dot{q}/\dot{q}=is-n>0$ on the $\dot{q}=0$ line. Hence q is increasing above the $\dot{q}=0$ line and decreasing below it as shown by the dynamic arrows in the diagram.

The discrepancy between the $\dot{T}=0$ and $\dot{q}=0$ lines implies that per capita interest receipts can decline when the economy's aggregate interest receipts are increasing, because the latter are not increasing at a rate high enough to offset the effect of population growth. Both lines intersect with the k axis at the same point because both lines satisfy the same condition $sf=\phi k$ at the intersection.

It must be noted also that both the $\dot{T}=0$ line and the $\dot{q}=0$ lines intersect with the k axis at the left-hand side of the $\dot{k}=0$ line. This can be proven by showing that the $\dot{T}=0$ line intersects with the $\dot{k}=0$ line below the k axis. The value for q on the k axis line is $1/s(\phi(n)k-sf(k))$. From the conditions that $s>n/i$, $f'<f(k)/k$, the following holds

$$\begin{aligned}\phi(n)k^*-sf(k^*)&<\phi(n)k^*-(n/i)f(k^*)\\&<nk^*(\phi(n)/n-r^*/i)\end{aligned}$$

However, $\phi(n)/n-r^*/i$ is negative (see below, Section 3).

The zero-income curve is also depicted by turning the production function upsidedown, i.e., the mathematical expression is $q=-f(k)$. It indicates a combination of the capital-labor ratios and per capita interest receipts which make the economy's income zero. Similar iso-income curves whose levels of disposable income are measured by intercepts at the q axis can be drawn by shifting the curves vertically.

3). That per capita income is always increasing along the arrows in the high saving ratio case can be proven as follows. Differentiating income with respect to time, we obtain, $\dot{y}=f'(k)\dot{k}+\dot{q}$, where $\dot{y}$, $\dot{k}$, and $\dot{q}$ are the time derivatives of the respective variables. If we substitute equations (9) and (10) for $\dot{k}$ and $\dot{q}$, we obtain $\dot{y}=f'(z-n)k+(is-n)q+(sf(k)-\phi\dot{z})$; $\dot{y}$ is an increasing function of q (payments) for a given k. We can say, therefore, that y is always increasing if $\dot{y}$ is positive when per capita interest payments have the lowest value, i.e., $q=-f(k)$. By substituting this relationship to the above equation, we get $\dot{y}=n(f-rk)+ikz(r/i-\phi/z)$. The first term is obviously positive. The second term is also positive if $\dot{K}>0$. The difference $(r/i-\phi/z)$ is net profit from the firms' point of view since r/i is the present discounted value of an increase in profit when real capital increases by one unit and ϕ/z is the cost of achieving that increase. In the private economy, net profit of investment must be positive for positive capital accumulation to take place. This can be more rigorously proven by using the convexity of the Penrose curve. For details, see Onitsuka, p. 95.

4). The $\dot{q}=0$ line in Figure 4 is positively sloped since

$$\left.\frac{dq}{dk}\right|_{\dot{q}=0}=-\frac{i}{is-n}\left(sf'-\frac{\phi}{r}f''k-\phi\right)>0$$

The system is stable with respect to q, for $\partial\dot{q}/\partial q=is-n<0$ on the $\dot{q}=0$ line.

5). Figure 5 is the same as Figure 4 except that the zero-capital flow line and the $\dot{q}=0$ line intersect with the k axis at the right-hand side of $\dot{k}=0$ due to the condition $s<\phi(n)k^*/f(k)$. This is so because the $\dot{q}=0$ line is positively sloped and intersects with the $\dot{k}=0$ line below the k axis; the zero-capital flow line also has to intersect with the k axis at the same point.

6). From equation (25), y^* is positive if

$$\frac{f(k^*)}{k^*}\frac{1}{i}-\frac{\phi(n)}{n}>0$$

since $(is-n)<0$ in the low-saving ratio case II. However,

$$\frac{f(k^*)}{k^*}\frac{1}{i}-\frac{\phi(n)}{n}>\frac{r^*}{i}-\frac{\phi(n)}{n}>0$$

REFERENCES

A. Amano, "International Capital Movements and Economic Growth," *Kyklos*, Fias, 4, 1965, *18*, 693–99.

G. H. Borts, "A Theory of Long-Run International Capital Movements," *J. Polit. Econ.*, Aug. 1964, *72*, 341–59.

———, "Growth and Capital Movements Among U.S. Regions in the Postwar Period," in J. F. Kain and J. R. Meyer, eds., *Regional Economics*, Cambridge 1971.

N. S. Buchanan, *International Investment and Domestic Welfare*, New York 1945, 166–80, 206–39.

J. E. Cairnes, *Some Leading Principles of Political Economy Newly Expounded*, New York 1900.

E. D. Domar, "The Effect of Foreign Investment on the Balance of Payments," in his *Essays in the Theory of Economic Growth*, New York 1957, 129–53.

K. Hamada, "Economic Growth and Long-Run Capital Movements," *Yale Econ. Essays*, spring 1966, *6*, 49–96.

J. A. Hanson and P. A. Neher, "The Neoclassical Theorem Once Again: Closed and Open Economies," *Amer. Econ. Rev.*, Sept. 1967, *57*, 869–78.

F. R. Harrod, *Toward a Dynamic Economics*, London 1948.

R. Hinshaw, "Foreign Investment and American Employment," *Amer. Econ. Rev. Proc.*, May 1946, *36*, 661–71.

C. Iversen, *Aspects of the Theory of International Capital Movements*, New York 1936, 19–30.

H. G. Johnson, "Equilibrium Growth in an International Growth," in his *International Trade and Economic Growth*, Cambridge 1965, 120–49.

M. C. Kemp, *The Pure Theory of International Trade*, New Jersey 1964.

C. P. Kindleberger, *International Economics*, Homewood 1968, 483–86.

H. B. Lary, "The Domestic Effects of Foreign Investments," *Amer. Econ. Rev. Proc.*, May 1946, *36*, 672–86.

A. P. Lerner, "On the Marginal Product of Capital and Marginal Efficiency of Investment," *J. Polit. Econ.*, Feb. 1953, *61*, 1–14.

R. E. Lucas, Jr., "Adjustment Costs and the Theory of Supply," *J. Polit. Econ.*, Aug. 1967, *75*, 321–34.

G. D. A. MacDougall, "The Benefits and Costs of Private Investment from Abroad: A Theoretical Approach," *Econ. Rec.*, Mar. 1960, *36*, 13–35.

A. Marshall, *Money, Credit, and Commerce*, London 1923.

J. E. Meade, *The Theory of International Economic Policy*, London 1955, vols. I and II.

L. A. Metzler, "The Process of International Adjustment Under Condition of Full Employment: A Keynesian View," in R. E. Caves and H. G. Johnson eds., *Readings in International Economics*, Homewood 1968, 465–86.

R. A. Mundell, "International Trade and Factor Mobility," *Amer. Econ. Rev.*, June 1957, *47*, 321–25.

———, "Capital Mobility and Stabilization Policy under Fixed and Flexible Exchange Rates," *Can. J. Econ.*, Nov. 1963, *29*, 475–85.

P. A. Neher, "International Capital Movements Along Growth Paths," *Econ. Rec.*, Sept. 1970, *46*, 393–401.

B. Ohlin, *Interregional and International Trade*, Cambridge, Mass. 1933.

Y. Onitsuka, "International Capital Movements and Patterns of Economic Growth," unpublished doctoral dissertation, Chicago Univ. 1970.

R. Solow, "A Contribution of the Theory of Economic Growth," *Quart. J. Econ.*, Feb. 1956, *70*, 65–94.

T. W. Swan, "Economic Growth and Capital Accumulation," *Econ. Rec.*, Nov. 1956, *32*, 334–61.

A. B. Treadway, "On Rational Entrepreneurial Behavior and the Demand for Investment," *Rev. Econ. Stud.*, Apr. 1969, *36*, 227–39.

H. Uzawa, "Marginal Efficiency of Investment and International Capital Movements," unpublished paper, Chicago Univ. 1968.

———, 'Time Preference and the Penrose Effect in a Two Class Model of Economic Growth," *J. Polit. Econ.*, July/Aug. 1969, *77*, 628–52.

The Demand and Supply of Durable Goods and Durability

By RICHARD W. PARKS*

There has been considerable interest recently in the question, will the durability of a good differ if it is supplied by a competitive industry rather than a monopolist. Until Peter Swan's recent paper the prevailing opinion was that a monopolist would restrict both the output and the durability of his product. (See David Martin, E. Kleiman and T. Ophir, R. Avinger, and David Levhari and T. N. Srinivasan.) Swan's analysis shows persuasively under the same essential assumptions made by the previous authors, that the market structure will not influence the choice of durability.

The unifying features of Swan's argument include the assumptions that consumers demand the services of durable goods, and that the *services* of assets of different durabilities are identical. Thus a consumer is indifferent between assets of different durabilities provided that the service prices are identical. Suppliers are then faced only with the problem of supplying the level of services demanded at minimum cost and the decision does not depend on the market structure of the service or asset market. This argument is robust with respect to the pattern of depreciation assumed for the asset.[1]

This paper extends the analysis of durability in several directions. Section I presents a simple model of the demand for consumer durable goods and shows the simplifications necessary to obtain demand relations of the type used in previous studies. In Section II, transaction costs are introduced as an important feature in markets for durable goods, and the effect of transaction costs on the demand for durables is derived. Then, in Section III, the conditions for the optimal choice of durability and output level are derived, and it is shown that although optimal durability is positively related to the level of transaction costs, the choice does not depend on market structure.

The choice of durability may be related to other aspects of the cost of using the good in addition to the economic life. The case where durability is related to the operating cost of providing services is considered in Section IV, and the relationship between durability and parameters of the operating cost structure is developed. It is shown in Section V that only if the average operating cost is related to the total quantity of services demanded will there be a differential effect on the choice of durability by a monopolist as opposed to a competitive industry.

Finally, in Section VI, we consider an important problem that has received no attention in the previous studies of durability. Previous studies have all assumed that the market would determine a single optimal durability; whereas typically we observe many durabilities being offered simultaneously. The final section of the paper applies the transaction cost model

---

* University of Washington. I would like to thank J. Allan Hynes, John McGee, Walter Oi, and Arnold Zellner for patient listening and helpful comments at various stages of the research for this paper. Dennis Shoemaker gave considerable help with some of the numerical calculations. I am very grateful for research support from the National Science Foundation.

[1] Several other recent papers explore aspects of the robustness of these results. For example, E. Sieper and Peter Swan focus on the ownership of initial stocks of the asset. Richard Schmalensee considers the problem of maintenance.

together with an assumption of consumer diversity based on wage differences and shows that the market solution will involve the production of an array of assets of different durabilities. We also show that a monopolistic supplier will generally offer consumers a different array of durabilities than that provided by a competitive industry. As a result there is a welfare loss on the supply side as well as the usual loss on the demand side of the monopoly solution.

I. Simple Models of the Demand for Consumer Durables

Consider a consumer at a point in time, t=0, who must choose an optimum consumption and investment path for his indefinite future. Assume he anticipates a certain income stream $Y(t)$. The path of prices for a perishable consumer good, $p_c(t)$, and the price path for a durable consumer good, $p_d(t)$, are assumed to be known or "expected" with certainty. The consumer's decisions are made with the goal of maximizing an intertemporal utility function that for simplicity we write as[2]

$$(1) \qquad \int_0^\infty U(C(t), Q(t))e^{-\gamma t}dt$$

where $C(t)$ represents the consumption of the perishable good, $Q(t)$ represents the flow of services from the stock of consumer durables. If the flow is proportional to the stock, as we shall assume, we can represent both by $Q(t)$. The instantaneous utility function $U(C, Q)$ and the discount factor, γ, are assumed to be unchanged over time.

On the above assumptions the consumer maximizes (1) subject to a wealth constraint

$$(2) \qquad W = \int_0^\infty Y(t)e^{-rt}dt = \int_0^\infty [C(t)p_c(t) + I(t)p_d(t)]e^{-rt}dt$$

where r represents the interest rate (assumed constant), where W, the consumer's wealth is the present value of anticipated income, and the right-hand side, with $I(t)$ representing gross purchases of the durable, is the present value of all consumer expenditures.

We assume that the durable exhibits a one-horse shay pattern of depreciation, that is, it provides a constant flow of services for its known life time T, then becomes useless.[3] The stock of the durable asset (hence the flow of services) at any time t is

$$Q(t) = \int_{t-T}^{t} I(t)dt$$

the sum of the amounts put in place during the prior interval of length T.

Alternatively the gross investment can be decomposed into net additions to the stock and replacement, thus

$$(3) \qquad I(t) = \dot{Q}(t) + I(t-T) = \dot{Q}(t) + Q(t-T) + \dot{Q}(t-2T) + \ldots$$

Substituting (3) in (2) the wealth constraint can be rewritten as

$$(4) \qquad W = \int_0^\infty [C(t)p_c(t) + (\dot{Q}(t) + \dot{Q}(t-T) + \dot{Q}(t-2T) + \ldots)p_d(t)]e^{-rt}dt$$

---

[2] The utility function is very restrictive with respect to the consumer's intertemporal choices, but it provides a simple vehicle for present purposes. Laurits Christensen and Dale Jorgenson, W. Erwin Diewert (1969), and Parks (1969), have experimented with more general representations. Limited empirical results suggest the need for the general form in explaining observed behavior.

[3] Parks (1966) incorporates a constant evaporation depreciation assumption. Levhari and Srinivasan use one-horse shay depreciation; Kleiman and Ophir use a nonconstant service stream for a fixed lifetime; Swan uses one-horse shay, evaporation, and a combination that is a special case of that considered by Kleiman and Ophir. Finally, Sieper and Swan use a general monotonic depreciation pattern. Qualitative results are not sensitive to the form chosen. My choice here is a matter of convenience.

The backward sum in (3) and (4) extends backward as far as need be to start the history of the durable stock.

This optimizing problem can be solved easily by calculus of variations techniques. The Euler conditions give the constraint (4) plus

$$(5)\quad U_{C(t)}e^{-\gamma t} = \lambda p_c(t)e^{-rt}$$

$$\begin{aligned}(6)\quad U_{Q(t)}e^{-\gamma t} = \lambda e^{-rt}[&rp_d(t) + rp_d(t+T)e^{-rT} \\ &+ rp_d(t+2T)e^{-r2T} + \dots \\ &- \dot{p}_d(t) - \dot{p}_d(t+T)e^{-rT} \\ &- \dot{p}_d(t+2T)e^{-r2T} - \dots]\end{aligned}$$

These conditions characterize an internal optimum path of consumption for both perishables and durables.

The interpretation of the necessary conditions is straightforward. The marginal utility of consumption at any point in time for either type of good must be proportional to the appropriate forward (discounted) price.[4] The expression in brackets in (6) represents the spot rental or service price for a unit of the durable good at time t. Although the general expression is rather unwieldy, it takes on a simple and recognizable form under more restricted assumptions about price movements. Suppose that the consumer expects the price of new durables to be unchanged from t onward at a level $p_d(t)$. The service price then reduces to

$$\begin{aligned}(7)\quad \pi(t) &= rp_d(t)[1 + e^{-rT} + e^{-2rT} + \dots] \\ &= \frac{rp_d(t)}{1 - e^{-rT}}\end{aligned}$$

In this simplified form the implicit rental price is equal to the rate of interest multiplied by the present value of the cost of maintaining a unit of the durable in perpetuity. Inverted, (7) shows that the price of a new durable good is the present value of the services provided in the interval (t, t+T) where the services are valued at $\pi(t)$. The latter expression figures prominantly in the above cited studies.

Suppose alternatively that the consumer expects prices to rise exponentially at rate θ. The service price can then be written as[5]

$$\begin{aligned}(8)\quad \pi(t) &= p_d(t)(r - \theta) \\ &\quad \cdot [1 + e^{-(r-\theta)T} + e^{-2(r-\theta)T} + \dots] \\ &= \frac{p_d(t)(r - \theta)}{1 - e^{-(r-\theta)T}}\end{aligned}$$

Comparison of expression (8) with (7) shows that the service price will be reduced by the capital gains term provided we compare them at comparable real rates of interest.[6]

If we assume that the consumer expects prices and interest rates to remain at the same level forever, and if we assume $r=\gamma$, the consumption paths are easy to obtain. Assume initially that the consumer starts with no stocks of durables. The desired path of consumption for perishable goods and for the services of durables will be constant under these strong assumptions; the relative demand for the two commodities is determined by their relative prices; the absolute levels by the size of the consumer's wealth.[7] The consumer will adjust his durable stocks to the desired level immediately (borrowing if necessary), and he will maintain them by repurchase when the assets wear out.

---

[4] Use of a more general utility framework that avoids the strong additivity for preferences over time would give necessary conditions to which we could attach the same interpretation.

[5] Clearly $(r-\theta)$ must be positive for the series to converge as indicated. This will be the case when r represents the nominal interest rate. The difference will then be the real rate, which can be assumed positive.

[6] This rental expression is to be compared with the rental expression $\pi(t)=p_d(t)[r+\delta-\dot{p}_d(t)/p_d(t)]$ which arises in the commonly used case of evaporating depreciation (at rate δ). Both forms embody the interest, depreciation, and price change or capital gains aspects, although they appear in different forms.

[7] If $r<\gamma$ the desired consumption paths will decline over time; if $r>\gamma$ the desired consumption paths will increase over time.

If the consumer holds initial stocks of durables and if there are markets for these goods, then their value enters the consumer's wealth in addition to the present value of income flows.[8] If the desired stock is above the initial holding, then it will be immediately augmented to that level and maintained. If the desired stock is below the initial holding, then part of this holding will be sold to reduce the stock to the desired level.

In the present formulation of the model there are no impediments to the kind of instantaneous adjustments described above. If there are limitations on borrowing and lending, however, it may take time for the consumer to adjust to the desired level of stocks. In my 1966 paper, for example, I examined the case where no borrowing and lending is possible, and showed that in some cases the model will lead to a partial adjustment model of the kind used in many empirical studies of both producers' and consumers' durables. Levhari and Srinivasan and Swan also consider cases where there are output limitations which prevent instantaneous adjustment to the desired stock.

Anticipated price variation over time will, of course, produce a more complicated set of consumption paths that cannot easily be given a general representation. It is enough for our present purposes to note that the form of the demand for services used in the above cited studies is the constant services version implied by assuming constant expected prices and equality between the interest rate and the rate of time discount. We will retain the special case for the subsequent analysis, noting that it is easy to derive the classical demand results for the effect of changes in relative prices and wealth on the desired consumption paths. The interest rate and technical lifetime provide an additional source of variation, but their effects operate through their influence on the service price for durables in a straightforward way.[9]

II. The Demand for Durables in the Presence of Transaction Costs

There is some systematic evidence and considerable casual evidence that transaction costs (including search and information costs) are significant when durable goods are purchased or resold. For nondurables, the frequent purchase provides a basis for acquiring and maintaining information about product prices and characteristics. Mistakes are short-lived and easily corrected. The infrequent purchases of durables means that consumers must devote resources to gaining information about the product, including, for example, information about its durability. George Stigler, Phillip Nelson, and others provide evidence of activities devoted to search and other aspects of transactions costs. There has been mention of the role of transaction costs as a factor limiting resale markets for some durables by Henry

---

[8] Suppose at $t=0$, consumer has a stock $Q(0) = \int_{-T}^{0} I(t)dt$ of durable goods. We need to compute the value of the stock at the prices prevailing at $t=0$. We will find the stock of new assets that is equivalent in terms of the present value of its service flow to the initial stock. The equivalent stock can then be evaluated at $p_d(0)$. The present value of the service flow generated by one unit of new stock at $t=0$ is $\int_0^T e^{-rt}dt = (1-e^{-rT})/r$. The present value of the service flow generated by the stock $Q(0)$ is $\int_0^T e^{-rt}[Q(0) - \int_0^t I(t-T)dt]dt$. This value divided by $1-e^{-rT}/r$ gives the number of new durable service equivalents, which can be valued at the price of a new durable. (This device requires unchanged service price.) With markets this value can be included as part of wealth. As Sieper and Swan show, it was the incorrect treatment of initial stocks that led many previous authors to erroneous conclusions.

[9] From (7)

$$\frac{\partial \pi}{\partial r} = \frac{p}{(1-e^{-rT})^2}\left[1 - e^{-rT} - rTe^{-rT}\right] > 0$$

and

$$\frac{\partial \pi}{\partial T} = \frac{-rpTe^{-rT}}{(1-e^{-rT})^2} < 0$$

An increase in the interest rate raises the rental price, while an increase in durability lowers it.

Miller and Avinger, but transactions costs have not been systematically introduced into the analysis of durables markets. The costs involved may take a variety of forms. One ingredient likely to be significant is the time cost of the consumer engaged in shopping. There may also be direct money costs associated with registration or installation. The most satisfactory way of including the time costs might involve a model that explicitly incorporates the time activities of the consumer. In the treatments by Gary Becker and Gilbert Ghez, for example, the time inputs for the consumption activities may be interpreted to include the time spent shopping for as well as consuming the goods. It is relatively easy to modify the model developed in Section I to allow the introduction of both time and money elements of the transaction costs associated with the purchase of the durable good.

Let us assume that the consumer seeks to maximize the utility functional

$$\int_0^\infty U[L(\mathrm{t}), Q(\mathrm{t})]e^{-\gamma \mathrm{t}}d\mathrm{t} \tag{9}$$

subject to a time constraint $H = L + N + S$, where H denotes the total number of hours available, L represents leisure hours, N the number of labor hours, and S the transaction or shopping time. It will also be assumed that the consumer pays, in addition to the stated price, an additional sum, denoted B, that represents the sum of resources other than the consumer's time that is required for the purchase of the durable good. The neglect of the nondurable consumer good is of no importance to the following results.

In the budget constraint, income is no longer exogenous but is the product of the wage rate (exogenous) and the hours worked. Shopping or transaction time is assumed proportional to the size of the purchase made, i.e., $S = bI$. The budget constraint can then be written as

$$\int_0^\infty w(\mathrm{t})N(\mathrm{t})e^{-r\mathrm{t}}d\mathrm{t} = \int_0^\infty [P_d(\mathrm{t}) + B]I(\mathrm{t})e^{-r\mathrm{t}}d\mathrm{t} \tag{10}$$

We can substitute the expression for transaction time and substitute for N using the time constraint to obtain[10]

$$\begin{aligned}\int_0^\infty w(\mathrm{t})H(\mathrm{t})e^{-r\mathrm{t}}d\mathrm{t} &= \int_0^\infty \{w(\mathrm{t})L(\mathrm{t}) + [p_d(\mathrm{t}) + B \\ &\quad + bw(\mathrm{t})]I(\mathrm{t})\}e^{-r\mathrm{t}}d\mathrm{t}\end{aligned} \tag{11}$$

Comparison of this budget expression with (2) brings out the formal similarities that allow a direct translation of the simpler results into the present context. Wealth in (11), assuming no initial holdings of the durable asset, consists of the present value of the consumer's time. It is allocated between the consumption of leisure and purchases of the durable. The cost of durables purchases is now written to include the value of time spent in transactions, evaluated at the wage rate, plus any other direct transaction costs. The conditions for an optimum consumption path are given by the Euler conditions

$$U_L(\mathrm{t})e^{-\gamma \mathrm{t}} = \lambda w(\mathrm{t})e^{-r\mathrm{t}} \tag{12}$$

$$\begin{aligned}U_Q(\mathrm{t})e^{-\gamma \mathrm{t}} = {} & \\ & \lambda e^{-r\mathrm{t}}[rp_d(\mathrm{t}) + rB + rbw(\mathrm{t}) \\ & + rp_d(\mathrm{t} + \mathrm{T})e^{-r\mathrm{T}} + rBe^{-r\mathrm{T}} \\ & + rbw(\mathrm{t} + \mathrm{T})e^{-r\mathrm{T}} + \ldots \\ & - \dot{p}_d(\mathrm{t}) - \dot{p}_d(\mathrm{t} + \mathrm{T})e^{-r\mathrm{T}} - \ldots]\end{aligned} \tag{13}$$

---

[10] This expression is correct provided that I is nonnegative. If the consumer incurs transaction costs in selling the asset, then his net price for sales would be $p_d(\mathrm{t}) - B$. In the present analysis, however, there would never be sales of the asset except possibly at $\mathrm{t} = 0$ if initial holdings of the asset exceeded the optimum level. The selling cost would only enter the evaluation of initial wealth.

Considering the case of stationary price, wage, and interest rate expectations, (13) reduces to

$$U_Q(t)e^{-rt} = \lambda e^{-rt}\frac{r(p_d + B + bw)}{1 - e^{-rT}} \tag{14}$$

The demand for durable services will depend, as in Section I, on the rental price, which is now written as

$$\pi = \frac{r(p_d + B + bw)}{1 - e^{-rT}} \tag{15}$$

The implicit rental price that enters the consumer's demand decision incorporates the transaction costs, part of which depends on the wage rate. For given asset price and durability the existence of transaction costs in the purchase of durables clearly raises the implicit rental price of durable services and will result in fewer services being demanded.[11] The adjustment of the consumer from his initial position to his desired position will be analogous to that described in Section I, given the way transaction costs have been introduced.[12]

The purpose of the analysis in Sections I and II is to show the conditions under which we can regard the demand for the services of a durable asset as a function simply of the current service price as seen by the consumer, that is, $Q=f(\pi)$ with $f'<0$. Transaction costs enter as a kind of tax on the purchases of the durable, and this "tax" has the feature that it depends on the durability of the asset. The tax only influences the consumer's decisions, however, by increasing the service price.

The demand function will also depend on the tastes, the wage rate, and other characteristics of the consumer. When the different characteristics have no direct influence on the service price of the asset, we can aggregate over consumers and obtain a market demand function of essentially similar form. As we shall see in a later section, however, consumer diversity introduced, for example, through differences in wage rates will directly affect the consumer's implicit service price for the asset. In this case, as we shall see in the final section, the market solution differs in several important ways from the results presented next.

The following section assumes that we can aggregate in a simple way and it derives the producer's profit maximizing conditions and shows how the transaction cost tax influences the producer's choice of durability and output.

III. The Output and Durability Decisions of Producers in a Market with Transactions Costs

In analyzing the supply decisions of the producers of the durable goods, it is convenient to maintain the cost assumptions of Swan, Levhari and Srinivasan, and others. Let there be a large number of potential producing units each with identical costs. These producing units can be operated either as independent competitors or they can be operated by a monopolist with no effect on costs. The long-run average (=marginal) cost of a unit of the asset depends only on the degree of durability chosen by the producer; that is, we can represent the cost per unit of an asset of given durability T by $D(T)$ where we assume $D'(T)>0$.

Consider the producer's decisions in a case where there are no existing stocks of the good, and where he plans to sell at the same price over time. The present value of net receipts can then be given by

---

[11] We assume noninferiority of durable services.

[12] I have assumed that the total resource cost per unit of the durable rises with the wage and at the constant rate b. These assumptions clearly abstract from many of the complexities of a search and transaction cost model. In fact if shopping time is only associated with search for lower price, then the effect of increased w on the total resource cost of the good may depend on the income elasticity and may be negative. The effect of wage change on durability (see Sections II and VI below) would then depend on the sign of dB^*/dw.

$$(16)\quad V = [Qp_d - QD(\mathrm{T})] \cdot [1 + e^{-r\mathrm{T}} + e^{-2r\mathrm{T}} + \ldots] = [Qp_d - QD(\mathrm{T})]\frac{1}{1 - e^{-r\mathrm{T}}}$$

where Q represents the level of sales of the asset needed to maintain the service flow at level Q. Expression (16) recognizes the fact, derived in Sections I and II, that sales will be periodic; consumers will make an initial purchase $Q=I(0)$ to produce the desired flow of services, Q, then will replace the stock as it wears out at intervals T.

Solving for p_d in the expression for the service price, $\pi=r(p_d+B^*)/(1-e^{-r\mathrm{T}})$ implied by (12), then substituting for p_d in (16) we obtain

$$(17)\quad V = \left[\frac{\pi(1 - e^{-r\mathrm{T}})}{r} - B^* - D(\mathrm{T})\right] \cdot \left[\frac{Q}{1 - e^{-r\mathrm{T}}}\right]$$

where $B^*=B+bw$ represents the total transaction cost per unit. Finally, let $\pi=g(Q)$ represent the inverse of the demand relation for services in terms of the service price that was derived in Section II. The Q of (17) represents both the desired level of services and the size of the periodic purchases required to maintain those services. The expression for the present value of net returns for the producer is then

$$(18)\quad V = \frac{Qg(Q)}{r} - \frac{B^*Q}{1 - e^{-r\mathrm{T}}} - \frac{QD(\mathrm{T})}{1 - e^{-r\mathrm{T}}}$$

In (18) the first two terms on the right represent the present value of revenue, and the third represents the present value of costs associated with maintaining a given service level Q with goods of durability T.[13] The monopolist will choose the combination of Q and T that maximizes V. The conditions for a maximum are

$$(19)\quad V_Q = \frac{\partial V}{\partial Q} = \frac{Qg'(Q) + g(Q)}{r} - \frac{B^* + D(\mathrm{T})}{1 - e^{-r\mathrm{T}}} = 0$$

$$(20)\quad V_\mathrm{T} = \frac{\partial V}{\partial \mathrm{T}} = \frac{-Q}{(1 - e^{-r\mathrm{T}})^2} \cdot [(1 - e^{-r\mathrm{T}})D'(\mathrm{T}) - (B^* + D(\mathrm{T}))re^{-r\mathrm{T}}] = 0$$

Equation (19) is essentially the condition that marginal revenue equals marginal cost, conditional on the durability selected. Equation (20), which can be written as

$$(21)\quad \frac{D'(\mathrm{T})}{B^* + D(\mathrm{T})} = \frac{r}{e^{r\mathrm{T}} - 1}$$

implicitly gives the desired durability T. Equation (21) is independent of Q hence also independent of the demand conditions for the services of the durable; thus we obtain the result, analogous to Swan's, that the degree of durability selected will be independent of the market conditions under which the good is sold. Output will, of course, be restricted in the case of the monopolist. We see from (18) that the transaction cost term can be viewed as an aspect of the producer's cost and that it is internalized by producers in deciding on the optimum durability.

We next show the following result: The lifetime or durability of the asset supplied will be a positive function of the transaction costs. Differentiating (20) with respect to B gives

---

[13] As in the case of a tax, the transaction cost tax can be viewed symmetrically as a reduction in average revenue or as an addition to cost.

$$(22)\quad \{(1-e^{-r\mathrm{T}})D''(\mathrm{T}) + [B^* + D(\mathrm{T})]r^2e^{-r\mathrm{T}}\}\,\frac{d\mathrm{T}}{dB^*} = re^{-r\mathrm{T}}$$

The expression on the right is positive and the factor in brackets must also be positive to satisfy the second-order conditions;[14] hence $d\mathrm{T}/dB^*$ is positive. Transaction costs can be viewed as a kind of tax on the services of the durable. An increase in transaction costs increases the service price as seen by the consumer and reduces the quantity of services demanded. This can be shown directly by differentiating (19) with respect to B^*.

$$(23)\quad V_{QQ}\frac{dQ}{dB^*} + V_{Q\mathrm{T}}\frac{\partial \mathrm{T}}{\partial B^*} = \frac{1}{1-e^{-r\mathrm{T}}}$$

V_{QQ} is negative, $V_{Q\mathrm{T}}=0$, hence dQ/dB^* is negative. Although the effect of increased transaction costs is partially offset by an increase in durability, the net effect is still to reduce the quantity of services.

IV. Operating Costs and the Choice of Durability

The choice of durability may affect certain direct operating costs as well as determine the lifetime of the asset. Consider the standard example of a lightbulb. It is possible to produce bulbs with different durability, but a major feature of the cost of providing lighting services, the power cost, tends to be positively related to the durability of the bulb. Longer-lived bulbs may use heavier filaments which use more electricity per lumen-hour.

Let the consumer derive utility from light services that are produced by combining a unit of light bulb service with a fixed amount of electricity. The amount of electricity required per light hour, $a(\mathrm{T})$ depends directly on the durability of the bulb used, i.e., $a'(\mathrm{T})>0$.[15] If prices of the bulbs and power together with the interest rate are expected to be unchanged over time, the consumer will have a stable demand for light services that will be a function of the service price which we can write as

$$(24)\quad \pi = \frac{rp_d}{1-e^{-r\mathrm{T}}} + p_v\cdot a(\mathrm{T})$$

where as before π is the service price, p_d, the price of a new asset (bulb), p_v is the price per unit of the variable input (electricity), and $a(\mathrm{T})$ is the technical parameter indicating the dependence of the variable input requirement on the durability selected. As before, let $\pi=g(Q)$ be the inverse demand relation between the service price and the quantity of services demanded.

Under the same assumptions about producer behavior and costs, the producer will try to maximize the present value of sales less costs, i.e.,

$$(25)\quad V = Q\left[\frac{g(Q) - p_v\cdot a(\mathrm{T})}{r}\right] - \frac{D(\mathrm{T})Q}{1-e^{-r\mathrm{T}}}$$

First-order conditions for a maximum are

$$(26)\quad V_Q = \frac{g(Q)+Qg'(Q)}{r} - \frac{p_v\cdot a(\mathrm{T})}{r} - \frac{D(\mathrm{T})}{1-e^{-r\mathrm{T}}} = 0$$

---

[14] The Hessian matrix is diagonal with elements

$$V_{QQ} = \frac{2g'(Q)+Qg''(Q)}{r}$$

and

$$V_{\mathrm{TT}} = \frac{-Q}{(1-e^{-r\mathrm{T}})^2}\cdot\{(1-e^{-r\mathrm{T}})D''(\mathrm{T}) + (B^*+D(\mathrm{T}))r^2e^{-r\mathrm{T}}\}$$

both of which must be negative; $V_{Q\mathrm{T}}=0$. The negativity of V_{QQ} corresponds to the familiar condition that when marginal cost is constant marginal revenue must be decreasing; V_{TT} is harder to characterize.

[15] It is easy to think of alternative examples in which larger T is associated with reduced use cost, i.e., $a'(\mathrm{T})<0$.

$$(27)\quad V_T = -\frac{p_v a'(T)}{r} - \frac{(1-e^{-rT})D'(T) - D(T)re^{-rT}}{(1-e^{-rT})^2}$$

Condition (26) is again the equality of marginal revenue and marginal cost, for a given durability T, where marginal revenue excludes the marginal cost of the variable factor. As in the case of the transaction costs, the variable cost can be viewed as a kind of tax on the use of the durable good, and as such it influences the firm's choice of output. Condition (27) is independent of Q; hence the choice of durability again is purely a cost minimizing consideration for the firm. The effect of durability on the operating costs is considered, but the choice will not be influenced by the market structure in which the firm operates. Only condition (26) would be altered in the case of a competitive firm, which would consider only average revenue $g(Q)$.

We can derive two results of interest for this model: When $a'(T) > 0$ an increase in the price of the related factor or an upward shift in $a(T)$ will increase the service price and reduce the quantity demanded of both the service and the durable good used to provide it; it will also lead to a decrease in the durability of the asset.[16]

V. A Case of Quantity-Durability Interaction

All of the problems that we have considered concerning the choice of durability share a common feature. The consumer is indifferent to the durability of the asset, because the services from assets of different durability are assumed perfect substitutes. The profit function in all of these cases has a structure in which marginal revenue is possibly a function of quantity but not durability, and marginal cost is a function of durability but not quantity. It is this separability that is responsible for the result that market structure, which determines only the shape of the demand function, does not effect the choice of durability.

The robustness of this result in the cases we have considered suggests several questions. What kinds of assumptions lead to the loss of the important separability? How general are these results for qualities of goods other than durability? Can we characterize the conditions under which the results will hold or fail to hold for characteristics of goods in general?

On the production cost side, it is clear that if marginal cost varies with output, the separability will be lost and the fact that monopoly and competition have different output levels would lead in general to different choice of durability. In the problem involving operating costs, one small change leads to a change in the results and suggests certain general implications. Without changing the demand features of the model, suppose that in the case of the light bulbs the quantity of light services used has an effect on the price of the operating input, electricity. We need to assume that consumers do not recognize this change (otherwise their demand decisions would be complicated) but that the effect is apparent at the market level. The monopolistic producer would have to take the effect into account and in general we would no longer expect the outcome to be independent of the market structure.[17]

---

[16] Formally, we find that again the Hessian has a simple diagonal structure, with negative elements required for a maximum. The comparative statics derivation gives

$$V_{QQ}\frac{dQ}{dp} = -V_{Qp} = \frac{a(T)}{r}, \qquad \text{hence } \frac{dQ}{dp} < 0$$

$$V_{TT}\frac{dT}{dp} = -V_{Tp} = \frac{Qa'(T)}{r}, \qquad \text{hence } \frac{dT}{dp} < 0$$

In cases where $a'(T) < 0$, conclusion (2) is reversed.

[17] In conditions (26) and (27), making p_v a function of Q, i.e., $p_v(Q)$ makes the two conditions fully simultaneous.

Finally, if we start from a framework in which the utility function of the consumer has as its arguments both a quantity and quality aspect of a good, a characteristic approach of the sort developed by Kelvin Lancaster, the demand function of the individual will have arguments that include both the price of the good and an index of the amount of the qualitative attribute per unit of the good. The revenue function as viewed by firms will in general involve quantity and quality choices that are interdependent. We would then expect differences in market structure to influence both choices.

VI. The Problem of Variety in the Demand and Supply of Durability

One implication of the previous models is particularly at odds with reality, and should be examined further. They imply that only a single durability will be provided by the market, whereas typically we observe variety rather than uniformity with respect to this choice. Until now, the analysis in this paper has followed the earlier examples and shares their defect, but the introduction of transaction costs and operating costs provides a useful vehicle for considering the problem of product diversity.

In the earlier analysis there is nothing that distinguishes consumers from one another. Implicitly the individual demand relations have been aggregated into a market demand relation, assuming that the same implicit rental price influences each consumer's choice. Where there are transaction costs or operating costs it is natural to distinguish classes of consumers that differ with respect to these costs and to trace the implications of these differences into the market for durable goods.

We have already developed in Section II a demand model that incorporates transaction costs, with the consumer's time (valued at his marginal wage rate) constituting a major aspect of the transaction cost. In Section III, we derived a producers' decision for the choice of output and durability assuming that consumers were identical with respect to their value of time. We now want to develop the producers' decisions in a market where consumer diversity enters the problem through differences among consumers (or consumer groups) in their wage rates.[18]

As in Section III, we assume that an increase in the value of consumers' time increases transaction costs. Higher transaction costs imply an increase in durability but also a decrease in the quantity of services demanded (holding real wealth constant).

Now suppose that there are n consumers that can be distinguished according to their wage rate but whose demand behavior is otherwise identical.[19] That is, the quantity of services demanded by the ith consumer is given by

$$Q_i = f\left[\frac{rp_d + rbw_i}{1 - e^{-rT}}\right] \tag{28}$$

where $Q=f(\pi)$ is the demand function corresponding with $\pi=g(Q)$ and where we neglect the nonwage related aspect of transaction costs.

A. *Competitive Supply*

Let us consider first the behavior of a group of competitive suppliers facing the diverse consumers. Competition will force each producer to sell at a price equal to marginal cost. Hence implicitly each consumer faces a schedule of durabilities and corresponding prices, $p_d = D(\mathrm{T})$. Each consumer will seek to buy that durability that minimizes the implicit service price. The ith consumer's choice of T that minimizes the rental cost (15) with $B^* = bw_i$ must satisfy

---

[18] A similar development could be made using the differences among consumers in operating costs as the basis for generating variety, but the argument is essentially the same and is not given. The important feature to note is that variety must enter through the rental function. Since durability does not enter the utility function, differences in taste cannot be used.

[19] Instead of n distinct consumers we could introduce n consumer groups each weighted by the number of individuals making up the group.

$$(29)\quad \frac{d\pi}{d\mathrm{T}} = \frac{(1-e^{-r\mathrm{T}})rD'(\mathrm{T}) - [D(\mathrm{T})+bw_i]r^2e^{-r\mathrm{T}}}{(1-e^{-r\mathrm{T}})^2} = 0$$

or

$$\frac{D'(\mathrm{T})}{D(\mathrm{T}) + bw_i} = \frac{r}{e^{r\mathrm{T}} - 1}$$

This condition coincides with condition (21) for the producer's optimal choice of durability in a market with transaction costs, so that the consumers' and producers' decisions coincide. The durability choice of individual producers is in a sense indeterminate. Given our special assumptions of constant marginal cost depending only on durability, a given competitive producer may decide to specialize in the production of one of the n durabilities desired by consumers, but there is nothing to prevent him from producing the entire array of n durabilities. The point of the argument, however, is that a competitive market facing n distinct consumers or consumer groups will produce n distinct durabilities of the asset, and the durabilities produced will be those that minimize the service costs for each of the consumers in the market.[20]

B. *Monopoly Supply*

Now consider the solution that will emerge if the asset is supplied by a monopolist. The monopolist seeks to maximize a profit function that is the sum of profits from sales in each of the n submarkets. If the monopolist could costlessly achieve and maintain a separation among the distinct consumers, he would maximize profits by choosing the profit maximizing solution, conditions (19) and (20), within each submarket. The analysis of Section III would apply for each consumer or group and the monopolist would produce the same array of durabilities as did the competitive industry, but it would sell each consumer a smaller quantity at a higher price than in the competitive case. To obtain this result, however, we must assume that the monopolist can act as a discriminating monopolist with respect to the different consumers. Discrimination arises because in this case, where the monopolist can deal separately with each consumer, the amount by which price exceeds marginal cost is different for each consumer. At first glance it may seem that use of the notion of discrimination is inappropriate in a case where different goods are sold at different prices in different markets. We have assumed, however, that the services of the goods of different durabilities are homogeneous, and the profit maximizing solutions within each separate market would imply different service prices in the different markets.

Now suppose that the barriers separating the consumers into nontrading units are removed, so that consumers can observe that the supplier is offering an array of durabilities and prices. We can show that generally consumers will find it possible to buy different durabilities from the ones they were previously buying and that by switching to another price-durability combination achieve a lower service price. To prove this statement, we begin by considering the optimum solution for a monopolist that was derived in Section III above. Suppose that expressions (19) and

---

[20] It is important to recognize that this solution that involves complete "customization" of goods for consumers depends again on the assumption that cost per unit is unrelated to quantity. A more realistic analysis might recognize that particularly at small levels of output, unit costs could be declining. This feature would force consumers to consider the tradeoff between their desire for customized goods and the lower cost attainable by accepting some standardization.

(20) determine the optimal durability and quantity for an isolated submarket consisting of consumers with wage rate w_i, hence with transaction cost $B_i = bw_i$. The corresponding price at which the monopolist would sell the asset is

$$p_d = \frac{(1 - e^{-r\mathrm{T}})}{r} g(Q) - B_i$$

Next consider the solution that would emerge in a second submarket in which the consumer's wage differs by a small amount from that in the first market. The transaction costs in the two markets will differ by an amount $dB = bdw$, and we have already derived the effect of a change in B on the equilibrium durability and quantity. The results are given in (22) and (23). The corresponding change in the equilibrium price is given by equation (30).

$$\begin{aligned}(30)\quad \frac{dp_d}{dB} &= \frac{(1-e^{-r\mathrm{T}})}{r} g'(Q) \frac{dQ}{dB} \\ &\quad + \frac{g(Q)}{r} re^{-r\mathrm{T}} \frac{d\mathrm{T}}{dB} - 1 \\ &= \frac{g'(Q)}{2g'(Q)+Qg''(Q)} \\ &\quad + \frac{g(Q)re^{-2r\mathrm{T}}}{(1-e^{-r\mathrm{T}})D''(\mathrm{T})+[B_i+D(\mathrm{T})]r^2e^{-r\mathrm{T}}} \\ &\quad -1\end{aligned}$$

The sign of this expression is indeterminate, but it is the rental expression derived below that we shall want to evaluate.

Finally, we should like to determine the rental or service price that would result if the individual in the first market could choose the durability-price combination offered to the second market. Define the cross-rental function for consumer i as

$$(31)\quad \pi[B_i;\ \mathrm{T}(B),\ p(B)] = \frac{r[p(B) + B_i]}{1 - e^{-r\mathrm{T}(B)}}$$

This function gives the implicit service price when consumer i buys the good $\mathrm{T}(B)$ at price $p(B)$.[21] When $B = B_i$ we have the own rental, when $B \neq B_i$ we have the cross rental relevant when consumer i buys a good designed for another submarket.

The derivative of the cross-rental function for consumer i with respect to B, can be written

$$\begin{aligned}(32)\quad & \frac{\partial\pi}{\partial B}[B_i;\ \mathrm{T}(B),\ p(B)] \\ &= \frac{\partial\pi}{\partial P}\cdot\frac{\partial P}{\partial B} + \frac{\partial\pi}{\partial \mathrm{T}}\,\frac{\partial \mathrm{T}}{\partial B} \\ &= \frac{-r[g'(Q) + Qg''(Q)]}{(1 - e^{-r\mathrm{T}})[2g'(Q) + Qg''(Q)]}\end{aligned}$$

where we have used expressions (22), (30), and the relation $g(Q) = \pi$. When evaluated at $B = B_i$ this derivative gives information about the stability of the multimarket monopoly solution when there are no artificial barriers between the submarkets. The sign of the derivative is determined by the sign of $g' + Qg''$. When this expression is zero the cross-rental function has a minimum at B_i and consumer i cannot achieve a lower rental price by switching to another durability. If $g' + Qg'' > 0$ (which implies that the rental function is positively sloped at B_i), he would never switch to a higher durability (the good designed for a submarket in which $B > B_i$), but he would find it cheaper to switch to a lower durability. Conversely if $g' + Qg'' < 0$ the consumer would switch to higher but not lower durability.

The class of demand functions for which $g' + Qg'' = 0$ is the only general class for which no switching can occur. Integrating we find that this differential equation implies demand functions of the form $\pi = g_0 + g_1 \log_e Q$, with $g_0 > 0$ and $g_1 < 0$. For this rather narrow class of functions there is a natural market separation that would permit the discriminating monopoly solution to be stable. Conditions (19) and (20) would be satisfied in each submarket, and no consumer could reduce the service cost by switching to the durability offered to another submarket.

---

[21] $\mathrm{T}(B)$ and $p(B)$ satisfy the conditions (19) and (20) for the optimal choice of T and p in a submarket characterized by consumers with transaction cost B. Thus the cross-rental function for consumer i treats B as the variable and B_i as a parameter.

In all other cases, however, the discriminating monopoly solution would be unstable in the absence of barriers between the submarkets, and consumers would tend to switch in the direction indicated above.[22]

These results imply that unless the markets are isolated from each other the profit maximizing solution for a monopolist selling in more than one market must generally incorporate additional constraints on the prices. For the case of two distinct submarkets the problem can be formulated as follows. The producer seeks to maximize the joint profit function

$$(33)\quad V(T_1, T_2, p_1, p_2) = \frac{p_1 - D(T_1)}{1 - e^{-rT_1}} f\left[\frac{r(p_1 + B_1)}{1 - e^{-rT_1}}\right] + \frac{p_2 - D(T_2)}{1 - e^{-rT_2}} f\left[\frac{r(p_2 + B_2)}{1 - e^{-rT_2}}\right]$$

with respect to the choice of durability and asset price in each of the two submarkets. For the solution to be stable we need, in addition to the usual second-order conditions, the inequality constraints

$$(34)\quad \frac{p_1 + B_1}{1 - e^{-rT_1}} \leqq \frac{p_2 + B_1}{1 - e^{-rT_2}} \qquad \text{or } \pi_{11} \leqq \pi_{12}$$

$$(35)\quad \frac{p_2 + B_2}{1 - e^{-rT_2}} \leqq \frac{p_1 + B_2}{1 - e^{-rT_2}} \qquad \text{or } \pi_{22} \leqq \pi_{21}$$

That is, the own rental must be less than or equal to the cross rental for both submarkets.

Consider the case where $B_1 < B_2$. If the demand function has the form $\pi = g_0 - g_1 \log_e Q$, then the maximization of V is found using conditions (19) and (20) in each submarket. The cross rental constraints will be satisfied automatically and they will be nonbinding. This is the case of naturally separated markets discussed above.

Both constraints (34) and (35) can never be binding simultaneously because written as equalities they are inconsistent. From the discussion of switching we also see that only the constraint preventing *one* of the switches will be relevant. If $g' + Qg'' < 0$, so that consumer 1 would tend to switch toward the more durable good T_2, it is the constraint $\pi_{11} \leqq \pi_{12}$ that is relevant and binding. Similarly, when $g' + Qg'' > 0$ the constraint $\pi_{22} \leqq \pi_{21}$ must be binding to prevent switching downward by consumer 2.

The symmetry of the problem allows us to analyze only one of the alternatives. Suppose the first constraint (34) is binding. We can write it as an equality and substitute in (33) to eliminate p_1. The resulting first-order conditions for a maximum are given in equations (36), (37), and (38). Condition (36) indicates that the choice of T_1 coincides with the choice in the competitive and isolated market monopoly case (compare with (20)); and the choice of T_1 is unaffected by the choice of T_2 and the prices. Condition (37) shows that the choice of T_2 will differ from the choice in the competitive and discriminating monopoly cases; because although the first term of (37) is analogous to (20), the presence of the second term will modify the solution. Taking account of the constraint needed to block a switch for a market 1 to

---

[22] With arbitrary demand functions it would be possible to have the condition $g' + Qg'' = 0$ satisfied at a particular point but not at others. Market separation would then be "accidental." I have ignored these possibilities in favor of the general demand classification given above.

$$\frac{\partial V}{\partial \mathrm{T}_1} = -f(\pi_{12}) \cdot \frac{(1 - e^{-r\mathrm{T}_1})D'(\mathrm{T}_1) - [B_1 + D(\mathrm{T}_1)]re^{-r\mathrm{T}_1}}{(1 - e^{-r\mathrm{T}_1})^2} = 0 \tag{36}$$

$$\begin{aligned}\frac{\partial V}{\partial \mathrm{T}_2} &= -f(\pi_{22}) \cdot \frac{(1 - e^{-r\mathrm{T}_2})D'(\mathrm{T}_2) - [B_2 + D(\mathrm{T}_2)]re^{-r\mathrm{T}_2}}{(1 - e^{-r\mathrm{T}_2})^2} + \frac{(B_1 - B_2)re^{-r\mathrm{T}_2}}{(1 - e^{-r\mathrm{T}_2})^2} \\ &\quad \cdot \left\{\frac{r[p_2 - D(\mathrm{T}_2)]}{(1 - e^{-r\mathrm{T}_2})} f'(\pi_{22}) + f(\pi_{22})\right\} = 0\end{aligned} \tag{37}$$

$$\begin{aligned}\frac{\partial V}{\partial p_2} &= \frac{r(p_2 + B_1)}{(1 - e^{-r\mathrm{T}_2})^2} f'(\pi_{12}) + \frac{f(\pi_{12})}{1 - e^{r\mathrm{T}_2}} - \frac{[B_1 + D(\mathrm{T}_1)]rf'(\pi_{12})}{(1 - e^{-r\mathrm{T}_1})(1 - e^{r\mathrm{T}_2})} \\ &\quad + \frac{r[p_2 - D(\mathrm{T}_2)]f'(\pi_{22})}{(1 - e^{-r\mathrm{T}_2})^2} + \frac{f(\pi_{22})}{(1 - e^{-r\mathrm{T}_2})} = 0\end{aligned} \tag{38}$$

market 2 leaves T_1 unchanged but results in larger (more durable) T_2.[23]

If instead the producer faced a demand function such that $g'+Qg''>0$, consumers in the high wage, high durability second market would want to switch to the lower durability good in the first market when confronted with the discriminating monopoly choice of prices and durability. To block the switch would require the imposition of constraint (34). The optimality conditions are analogous to those in (36)–(38) but with the subscripts reversed. In this case the higher durability T_2 would be unchanged from the discriminating monopoly choice, but the lower durability T_1 would be reduced.[24]

We may summarize the propositions proved in this section for the situation where there are two consumer markets characterized by different wage rates (or different variable costs).

Proposition 1. Under competitive supply, two durabilities will be offered. In each market the choice of durability is that which minimizes the rental or service cost. The price of the asset equals marginal cost in each case.

Proposition 2. Under monopoly supply when the two markets are costlessly isolated from one another, the producer will offer the same array of durabilities as under competition, but in each market will sell a smaller quantity at a price above marginal cost.

Proposition 3. The discriminating monopoly solution of Proposition 2 will generally be unstable if the markets communicate and if consumers in both markets have information about durability and price for both goods. If the demand function $\pi = g(Q)$ has the property $g'+Qg''<0$, (case 1), consumers in the low durability market will incur lower service cost if they buy the higher durability good. If $g'+Qg''>0$, (case 2), consumers in the high durability market will find it cheaper to switch to the lower durability. Only if $g'+Qg''=0$ will the markets exhibit a natural separation in the sense that neither consumer finds it to his advantage to switch to another durability. Only in this case will the discriminating monopoly solution of Proposition 2 be stable.

---

[23] In (37) the expression in brackets in the second term is the condition marginal revenue minus marginal cost for market 2. It is equal to zero for the isolated market choice of p_2 and T_2. (Condition (20) and the price analog of (19).) But we have seen that for the isolated market choice of p_1, T_1, and p_2, T_2 the rental inequality (33) is violated. (We are still in a context where $g'+Qg''<0$.) In order to satisfy the constraint and prevent switching, p_2 must rise and p_1 fall (T_1 and T_2 initially held constant.) But a rise in p_2 must reduce marginal revenue, hence with marginal cost constant, the expression in brackets becomes negative. Together with $B_1 > B_2$ this change means that the second term is positive. The adjustment in T_2 must be upward to make the entire expression zero.

[24] The argument parallels that in fn. 23—but in this case p_1 rises and p_2 falls.

Proposition 4. When the discriminating monopoly solution is unstable because of communication, profits are maximized by imposing a constraint that requires the own rental price to be less than or equal to the cross-rental price for each consumer. Only one of these constraints will be binding, the one required to block the switching. The constrained monopoly solution will involve a different array of durabilities from that offered under competition or under a discriminating monopolist selling in isolated markets. In case 1, the lower durability is unchanged but the higher durability is increased. The corresponding prices p_1 and p_2 are lower and higher, respectively, than their counterparts under the discriminating monopoly solution. In case 2 the higher durability is unchanged, but the lower durability is reduced, and p_1 and p_2 are higher and lower respectively than their counterparts.

Proposition 5. Since the imposition of the constraint $T_1 = T_2$ would be an additional binding constraint, the le Chatelier principle implies that profits would be reduced. This implies that a two-durability solution dominates a single-good solution.

Each of these propositions is illustrated numerically in the Appendix. Propositions 1, 2, 3, and 5 generalize easily for situations where the number of markets exceeds two. I have not found a convenient way to express Proposition 4 in the general case, however, because the number of constraints required to ensure that the own rental price does not exceed any of the cross-rental prices grows rapidly with the number of markets. This seems to be a mechanical rather than logical problem and the general result that the array of durabilities is affected should follow. One other generalization that is desirable but not given would allow different demand functions in the different markets.

VII. Conclusion

This paper has examined the determinants of the durability of a consumer asset. A number of previous studies have analyzed this problem but have disagreed on the question of whether the market structure would affect the durability provided by the market. In the case of a market comprised of homogeneous consumers, the results of this paper support the results of Swan in showing that market structure will not affect the durability provided. This paper extends his result to the case in which there are transaction costs associated with the *purchase* of the asset and to the case in which there are other costs associated with the *use* of the asset and where these use costs are related to durability. We can show the effect of a change in transaction costs or use costs on the level of durability provided.

Finally, the introduction of transaction or use costs provides a natural way to abandon the rigid assumption that the market provides only a single durability. When transaction costs are associated with the use of the consumer's time and when consumers differ according to their wage rates, we find that the analysis changes in important ways. The competitive market will provide a variety of durabilities, under the simple constant cost assumptions made here as many durabilities as there are distinct consumers. The monopolist will, in general, provide a different array of durabilities except when the distinct consumer groups fall into noncommunicating submarkets that essentially permit the monopolist to discriminate among customers. These cases are distinguished by the shape of the service demand function, and this shape also determines the way in which the durability array will be modified.

APPENDIX

Numerical Examples

In the process of deriving the results of Section VI, dealing with the variety of durable goods under alternative market situations, my intuition was aided substantially

TABLE 1—TWO MARKET DISCRIMINATING MONOPOLY SOLUTION (unstable upward)

Demand Function $\pi = 100 - .01Q$

Market	Wage	Q Quantity	T Durability	P Asset Price	Rental Matrix π_{ij} (consumer i buys durability j)		Profit
1	2.00	4811.72	6.32	241.25	51.88	51.31	2,315,273.14
2	10.00	4739.10	9.58	314.44	53.58	52.60	2,245,911.95
Total							4,561,185.09

by several numerical examples. In other cases numerical examples were used to test theoretical results that did not at first seem obvious. The examples that are presented in this Appendix serve to illustrate the basic comparative statics results of the model as well as many of the results on variety.

Using the notation from the body of the paper we assume for the first several examples that the basic demand function for services can be written in inverse form as the linear expression

$$\pi = g_0 + g_1 Q = 100 - .01Q \tag{A1}$$

The cost function, the marginal (≡average) cost for a unit of given durability T is assumed to be

$$D = d_0 + d_1 \mathrm{T} = 3 + 2\mathrm{T} \tag{A2}$$

Both π and D can be denominated in dollars, T is expressed in years.

The first exercise exhibits the basic comparative statics of a change in the consumer's wage rate from one consumer group to another and shows the instability of the discriminating monopoly solution.

Let there be two distinct consumers or consumer groups that are distinguished by their marginal wage rate. The wage rate for the first group, $w_1 = \$2$, the wage rate for the second group, $w_2 = \$10$. Both consumers (groups) are assumed to have a demand for services given by (A1) where π is computed to include the value of time spent in buying and replacing the asset. For simplicity the amount of time required per unit, b, is assumed to be unity; hence the transaction cost is just equal to the wage rate, w_i.

Following the exposition in Section VI, we assume that the producer can costlessly isolate the two submarkets and seeks to maximize the present value of net returns given costs expressed by (A2). Conditions (19) and (20), with $B^* = w_i$, give the quantity and durability that is optimal for each market. Table 1 gives the results of the calculations for the two markets.

We observe from the results that as the comparative statics predict, an increase in the transaction costs increases the optimal durability when comparing the low wage and the high wage markets and also increases the rental price and decreases the quantity of services. In this case the asset price also increases. In the absence of barriers separating the two markets, consumers in both groups would have the choice of two durabilities at their corresponding asset prices. From the rental price matrix in Table 1 it is clear that consumer 1 would not want to buy durability 1 because he could obtain a lower rental price by buying the good designed for market 2; that is, the cross-rental price, π_{12}, is less than the own rental price, π_{11}. Consumer 2, however, cannot lower his rental price by switching because as the rental price matrix shows, π_{22} is less than π_{21}. The final column of Table 1 shows the profits obtainable in the discriminating monopoly case where the markets can be costlessly isolated.

In the absence of specific barriers to cross-market purchasers, the producer who attempted to sell the two durabilities at the discriminating monopoly prices would find that he sold only goods of durability T_2. His profits when both consumer groups bought only durability T_2 would be reduced to \$4,553,299.30. The producer could *improve*

TABLE 2—PROFIT MAXIMIZING NONDISCRIMINATING MONOPOLY SOLUTION

Demand Function $\pi = 100 - .01Q$

Market	Wage	Quantity	Durability	Asset Price	Rental Matrix		Profit
1	2.00	4840.15	6.32	239.92	51.59	51.59	
2	10.00	4710.67	9.61	316.80	53.30	52.89	
Total							4,561,022.77

profits by optimally designing a *single* good that would take account of the structure of the combined market. That good would be a compromise good with durability 8.1961 selling for \$286.39, and profits from this policy would be increased to \$4,556,380.94. It was shown in Section VI, however, that this one-good solution is always dominated by a two-good solution that takes account of the tendency to switch and incorporates as a constraint a durability selection and pricing policy that maintains the two markets intact. The conditions for this solution are given by expressions (36), (37), and (38). In the present context these conditions imply the solution presented in Table 2. It is clear from the last column of Table 2 that profits from this solution although not as high as the discriminating monopoly solution, nevertheless dominate the single-good solution and in fact represent the highest profits attainable in the absence of some inexpensive mechanism to isolate the two markets. Notice that this profit-maximizing solution for the monopolist involves the same choice of durability for market 1 as occurred in the discriminating monopoly solution, but the choice in market 2 calls for a slightly higher durability. The price charged for good 1 is below the discriminating monopoly choice; whereas the price for good 2 is above the discriminating monopoly choice, although for a good of greater durability. Since the array of durabilities offered in a competitive market coincides with that determined in the discriminating monopoly solution, it is clear that, in general, when the markets cannot be isolated, a monopolist would offer a different array. Since the competitive and discriminating monopoly choices of durability have the property that they minimize the cost of providing services to each submarket, the monopolistic solution, in general, implies a social loss on the cost side in addition to the usual monopoly loss associated with the restriction of quantity.

Section VI derives the result that for one class of demand functions, namely demand functions that have the form, $\pi = g_0 + g_1 \log_e Q$ with $g_0 > 0$ and $g_1 < 0$, the markets would be naturally separated. The monopolistic producer could maximize profits by choosing the profit-maximizing price and durability within each submarket separately. Using the specific example, $\pi = 100 - 10 \log_e Q$, we illustrate the separation by considering three markets characterized by wages $w_1 = \$2$, $w_2 = \$5$, $w_3 = \$10$. The solutions obtained by maximizing profits separately within each submarket are given in Table 3.

The rental matrix shows that in each row the own rental price, π_{ii} is less than either of

TABLE 3—THREE MARKET DISCRIMINATING MONOPOLY SOLUTION (stable)

Demand Function $\pi = 100 - 10 \log_e Q$

Market	Wage	Quantity	Durability	Asset Price	Rental Matrix			Profit
1	2.00	5560.57	6.32	62.53	13.76	13.80	13.92	556,057.64
2	5.00	5240.90	7.78	72.67	14.40	14.35	14.40	524,090.52
3	10.00	4808.83	9.58	83.84	15.47	15.28	15.21	480,883.77

TABLE 4—TWO MARKET DISCRIMINATING MONOPOLY SOLUTION (unstable downward)

Demand Function $\pi = 20400 - 3000\ log_e\ Q + Q$

Market	Wage	Quantity	Durability	Asset Price	Rental Matrix		Profit
1	2.00	443.31	6.3271	12002.70	2560.44	2560.90	11,334,269.30
2	10.00	443.01	9.5895	15791.19	2562.15	2562.20	11,327,832.50
Total							22,662,101.80

the two cross-rental prices, π_{ij}, $i \neq j$. This implies that consumers in each of the three markets are content with the good and price designed for their market and cannot achieve lower service costs by switching to any of the other available goods. This class of demand functions forms a dividing class between the class of demand function that are either convex downward but with less curvature or concave (the linear demand function used above is a member), and the class of demand functions that are convex and have greater curvature. In the former class, illustrated by the linear demand function, consumers in lower wage markets will desire to shift to higher durabilities when confronted with the discriminating monopoly solution. In the second class, however, the reverse is true and consumers in the higher wage market will desire to shift to lower durabilities when confronted with the discriminating monopoly solution. This latter class of functions is characterized by demand curves $\pi = g_0 + g_1\ log_e\ Q + g_2 Q$ with the coefficients satisfying $g_2 > 0$ and $g_1 < -2g_2 Q$. We shall use as an example the demand function $\pi = 20400 - 3000\ log_e\ Q + Q$. For the two markets characterized by $w_1 = \$2.00$ and $w_2 = \$10.00$, the discriminating monopoly solution is given in Table 4.

The rental matrix again shows the direction in which consumers would want to switch in the absence of barriers separating the submarkets (in this case downward). The monopolist would have to take into account the added constraint on the rental prices that would prevent switching to achieve his maximum profit solution. The solution in this case is given in Table 5. In this situation we see that the durability for the higher wage market is unchanged but that the durability appropriate for the lower market has been reduced slightly.

This final example, like that given in Table 2, illustrates the general result that we can expect a monopolistic producer to offer a different array of durabilities than that produced under competition. Only in the case where the monopolist sells in noncommunicating markets or where the demand function has the property $g' + Qg'' = 0$ do the monopoly and competitive durabilities coincide. The array of durabilities offered may be modified in the direction of either greater or less durability and as the examples illustrate, the direction again depends on the shape of the demand function. In either case, however, the modification implies increased cost associated with the provision of the service stream from the durable good.

TABLE 5—PROFIT MAXIMIZING NONDISCRIMINATING MONOPOLY SOLUTION

Demand Function $\pi = 20400 - 3000\ log_e\ Q + Q$

Market	Wage	Quantity	Durability	Asset Price	Rental Matrix		Profit
1	2.00	443.31	6.3270	12002.70	2560.47	2560.88	
2	10.00	443.01	9.5895	15791.03	2562.17	2563.17	
Total							22,662,101.00

REFERENCES

R. L. Avinger, Jr., "The Economics of Durability," unpublished doctoral dissertation, Duke Univ. 1968.

G. S. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493–517.

L. R. Christensen and D. W. Jorgenson, "Intertemporal Optimization and the Explanation of Consumer Behavior," paper presented at the Winter Meeting of the Econometric Society, Dec. 1968.

W. E. Diewert, "Intertemporal Consumer Theory and the Demand for Durables," *Econometrica* forthcoming 1974.

G. Ghez, "Nonmarket Activities and Life Cycle Consumptions," unpublished doctoral dissertation, Columbia Univ. 1970.

E. Kleiman and T. Ophir, "The Durability of Durable Goods," *Rev. Econ. Stud.*, Apr. 1966, *33*, 165–78.

K. J. Lancaster, "A New Approach to Consumer Theory," *J. Polit. Econ.*, Apr. 1966, *74*, 132–57.

D. Levhari and T. N. Srinivasan, "Durability of Consumption Goods: Competition Versus Monopoly," *Amer. Econ. Rev.*, Mar. 1969, *59*, 102–07.

D. D. Martin, "Monopoly Power and the Durability of Durable Goods," *Southern Econ. J.*, Jan. 1962, *28*, 271–77.

H. L. Miller, Jr., "On the Theory of Demand for Consumer Durables," *Southern Econ. J.*, Apr. 1961, *27*, 298–304.

P. J. Nelson, "Information and Consumer Behavior," *J. Polit. Econ.*, Mar./Apr. 1970, *78*, 311–29.

R. W. Parks, "The Demand for Consumer's Durables: A Utility Maximization Approach," paper presented at the Winter Meeting of the Econometric Society, Dec. 1966.

———, "The Demand for Consumer Durables," unpublished paper, Univ. Chicago 1969.

R. Schmalensee, "Market Structure, Durability, and Maintenance Effort," unpublished paper 1972.

E. Sieper and P. Swan, "Monopoly and Competition in the Market for Durable Goods," *Rev. Econ. Stud.*, forthcoming.

G. Stigler, "The Economics of Information," *J. Polit. Econ.*, June 1961, *69*, 213–25.

P. L. Swan, "Durability of Consumption Goods," *Amer. Econ. Rev.*, Dec. 1970, *60*, 884–94.

Competitive Price Adjustments: An Empirical Study

By B. T. McCallum*

A striking contrast exists between the theoretical treatments of competitive price dynamics, and that usually found in empirical studies. Despite recent objections,[1] the theoretical literature on price dynamics is dominated by the law of supply and demand, i.e., the proposition that price changes depend upon excess demand. This dependence, furthermore, is usually reflected in a nondecreasing, continuous, homogeneous function. In the econometric literature, on the other hand, various price adjustment "mechanisms" have been utilized, most with rather casual theoretical justification.[2] The contrast exists, it would appear, largely because of two features of the theoretical expression of the law of supply and demand that make the latter almost nonoperational.[3] First the standard version, $dp/dt=\phi[E(p)]$, is expressed in terms of continuous time, whereas actual data necessarily refer to periods of discrete length. It is not obvious, furthermore, which (if any) of several plausible discrete versions should be taken to correspond to the standard formulation. Secondly the usual theoretical models fail to recognize holdings of inventories, despite the presumed influence of inventory levels on price movements in most nonservice industries.[4]

The aims of this paper, accordingly, are to formulate operational models of competitive price adjustment processes—about half incorporating inventory holdings—and to explore the empirical adequacy of each, using data on the highly atomistic *U.S.* lumber manufacturing industry. Section I introduces several inventory-free versions of the law of supply and demand, and discusses in general terms steps needed to render these operational. Section II develops a number of inventory models, incorporating alternative price expectation hypotheses, in which price changes reflect purposive dynamic behavior on the part of firms. Data sources are described in Section III, while empirical results are presented and discussed in Section IV.

I. Operational Laws of Supply and Demand

The theorists' standard version of the single-market law of supply and demand, namely

$$(1) \qquad \frac{dp}{dt} = \phi[D(p) - S(p)] = \phi[E(p)]$$
$$\phi' > 0, \phi(0) = 0$$

---

* Associate professor of economics, University of Virginia. I am indebted to Byron G. Spencer, Dov Paris, an anonymous referee, and the managing editor for helpful suggestions. The Wilson Gee Institute provided financial support.

[1] See, in particular, Donald Gordon and Allan Hynes. For an argument that the critics have gone too far, see my 1972 paper.

[2] Somewhat *ad hoc*, if plausible, "disequilibrium" or "demand" variables appear in price equations estimated by Charles Schultze and Joseph Tryon, Thomas Courchene, F. Rushdy and Phillip Lund, and Wesley Yordan. On the other hand, the admirably explicit models of Edwin Mills, pp. 82–136, and George Hay are not based on *competitive* theory.

[3] Of course it cannot be denied that some economists view the law as fundamentally inadequate—see Kenneth Arrow and, again, Gordon and Hynes. But such arguments have not been put forth by empirical workers.

[4] This presumption shows up clearly in the paper by Otto Eckstein and Gary Fromm, as well as in Courchene and in Schultze and Tryon.

refers to a continuous process, whereas the econometrician must work with data for discrete periods. If we let P_t and E_t denote the (average) price and excess demand during period t, then we obtain two "natural" discrete versions of (1). With the counterpart of ϕ taken to be linear, these are

$$P_t - P_{t-1} = kE_t \tag{2}$$

and

$$P_t - P_{t-1} = kE_{t-1} \tag{3}$$

where k is a positive constant. It is not clear which of these two specifications is superior. Even if one agrees that price changes can only be caused by excess demand occurring at some previous point of time, he is not necessarily led to the lagged version (3) for empirical purposes, for the period of observation may be considerably longer than the length of the time lag involved.[5] In such a case, specification (2) could be more appropriate than (3).

Indeed, the possibility that separate economic units react with different speeds suggests that E_t and E_{t-1} might both be important determinants of $P_t - P_{t-1}$, or even that an extended distributed lag formulation might be called for.[6] We are thus led to introduce

$$P_t - P_{t-1} = k[\theta E_t + (1-\theta)E_{t-1}] \qquad 0 < \theta < 1 \tag{4}$$

and

$$P_t - P_{t-1} = (1-\lambda)kE_t + \lambda(P_{t-1} - P_{t-2}) \qquad 0 < \lambda < 1 \tag{5}$$

as additional specifications worthy of exploration.[7]

TABLE 1—COEFFICIENT SIGNS IMPLIED BY THE VARIOUS MODELS

	Explanatory Variables							
Model 1	y_t	y_{t-1}	w_t	w_{t-1}	P_{t-1}	P_{t-2}	I_{t-1}	I_{t-2}
S&*D*-I	+		+		+			
S&*D*-II		+		+	?			
S&*D*-III	+	+	+	+	?			
S&*D*-IV	+		+		+	−		
INV-I	+		+		+		−	
INV-II	+		+		+	−	?	−
INV-III	+		+		+	?	−	
INV-IV	*		*		?		*	
INV-V	+		+				−	

Note: All coefficients labelled * should be of the same (indeterminate) sign.

Since excess demand cannot be directly observed, operational expressions must be based upon specifications of demand and supply functions for the commodity in question. For the moment, let these be given by[8]

$$D_t = \beta_1 P_t + \beta_2 y_t \qquad \beta_1 < 0, \beta_2 > 0 \tag{6}$$

and

$$Q_t = \gamma_1 P_t + \gamma_2 w_t \qquad \gamma_1 > 0, \gamma_2 < 0 \tag{7}$$

Then by definition the level of excess demand will be

$$E_t = (\beta_1 - \gamma_1)P_t + \beta_2 y_t - \gamma_2 w_t \tag{8}$$

Our procedure is to substitute expression (8) into equations (2)–(5) and solve in each case for P_t. The first of the resulting formulations, based on adjustment equation (2), can be written as

$$P_t = \pi_1 y_t + \pi_2 w_t + \pi_3 P_{t-1} \tag{9}$$

where the signs of the underlying parameters imply that $\pi_1, \pi_2, \pi_3 > 0$. These signs, as well as those implied in the other cases, are shown in Table 1. The models resulting from equations (2) through (5) are labelled *S&D*-I through *S&D*-IV, respectively.

---

[5] For a discussion of the issues see Ragnar Bentzel and Bent Hansen.

[6] Distributed lag versions of the law have been utilized, in more highly aggregated contexts, by Ray Fair and in my 1970 paper.

[7] Equation (5) is of course the Koyck version of the geometric distributed lag equation $P_t - P_{t-1} = k(1-\lambda)[E_t + \lambda E_{t-1} + \lambda^2 E_{t-2} + \ldots]$.

[8] Here we take y (income) and w (wages) as the sole nonprice determinants only for simplicity of exposition. In the empirical sections below additional variables are used.

It should be noted that these operational versions of the traditional theory do not treat "supply" as synonymous with "production." Indeed, our approach does not identify supply with any *directly* observable magnitude, but instead expresses it in terms of its determinants. This treatment is intended to be less restrictive than alternative possibilities, and thus to provide the traditional theory with a relatively generous opportunity for satisfactory empirical performance.[9] Our treatment retains, nevertheless, the crucial implication that explicit recognition of inventory levels is unnecessary for the explanation of price changes.[10]

It should be clear, incidentally, that while (9) has for convenience been written with P_t on the left-hand side, it describes price adjustments, as well as levels. One cannot predict values of P_t without simultaneously predicting values of ΔP_t.

Before turning to models incorporating inventory holdings, we should consider the possibility of explaining price changes with a "partial adjustment" model of the type introduced by Marc Nerlove, and used with considerable success in a wide variety of studies.[11] Application would involve solving (6) and (7) for the "equilibrium" price P_t^* that equates D_t and Q_t, and then substituting into the adjustment equation

$$(10) \quad P_t - P_{t-1} = \mu(P_t^* - P_{t-1}) \quad 0 < \mu < 1$$

This procedure, however, leads to

$$(11) \quad P_t = \frac{\mu\beta_2}{\gamma_1 - \beta_1} y_t - \frac{\mu\gamma_2}{\gamma_1 - \beta_1} w_t + (1 - \mu)P_{t-1}$$

which is operationally equivalent to (9). Thus there is no need to give separate consideration to this approach.

II. Inventory Models

Let us now consider several models that explicitly recognize holdings of finished goods inventories. In each case the underlying hypothesis is the same, i.e., that inventory levels are related to current and anticipated future prices as posited in the so-called "supply-of-storage" literature.[12] The implied relationship, together with those describing price expectation and production cost behavior, is used to generate a product supply function that takes into account inventory adjustments which approximate those of present-value maximizing firms. Price levels in all periods are those that equate the resulting industry supply with market demand. The different models result from alternative assumptions regarding price expectations.

The key relation in the supply-of-storage literature states that, for a present-value maximizing competitive firm, the optimal level of inventory to hold at the end of a period is that at which the marginal cost of storage equals the price change expected between the current and the succeeding period. Symbolically, this optimality condition can be written as

$$(12) \qquad P_{t+1}^e - P_t = \psi(I_t)$$

---

[9] An additional virtue of the approach is its applicability to nonstorable, as well as storable, commodities. In the former case, some parties will be dissatisfied—off of their demand or supply functions—when a nonequilibrium price prevails. But the *desired* demand and supply quantities will still be given by (6) and (7), and thus excess demand by (8).

[10] Thus our formulation of the theory does not imply that inventories are not held or that adjustments are not made, only that they need not be given explicit attention. Indeed, since price changes occur when disequilibrium prevails, such changes will tend to take place while firms are adjusting inventory levels, production plans, or both.

[11] See, for example, those edited by Arnold Harberger. Few investigators have used the partial adjustment model for *price* equations, but Robert Neild's well-known study can be so viewed; see my 1970 paper, p. 149.

[12] Leading examples include work by Michael Brennan, Lester Telser, and Holbrook Working.

where P^e_{t+1} is the price expected for period t+1, I_t is the level of inventory carried from t into t+1, and ψ is a marginal storage cost function with $\psi' > 0$ and $\psi(0) < 0$.[13] If corner solutions are assumed away, (12) represents a first-order condition for maximization of the present value of future net revenues, another such condition being that marginal production cost equals price. Michael Brennan (1959) has shown, furthermore, that similar conditions hold at the industry level, provided that P^e_{t+1} is the same for all firms. Thus if P^e_{t+1} were given, industry supply, i.e., production plus inventory drawn-down, would be a function of P_t, and the market clearing price would be that which equated supply (so defined) to demand.

In practice, P^e_{t+1} is not given, so an operational version of the theory requires some assumption regarding price expectations. All in all we shall explore five different possibilities, but for the moment let us focus attention on just one—the familiar adaptive expectations formula,

$$P^e_{t+1} - P^e_t = \delta(P_t - P^e_t) \tag{13}$$

where $0 < \delta < 1$. Let us also assume that ψ can be adequately approximated with a linear expression,

$$\psi(I_t) = \alpha_0 + \alpha_1 I_t \tag{14}$$

where $\alpha_0 < 0$ and $\alpha_1 > 0$. Finally, to make the present model comparable to those of Section I, let us again use (6) and (7) for D_t and Q_t, keeping in mind that the latter now represents production rather than supply. Under these assumptions, the market-clearing condition $D_t = Q_t + I_{t-1} - I_t$ can be written as

$$\begin{aligned}\beta_1 P_t + \beta_2 y_t = \gamma_1 P_t + \gamma_2 w_t + I_{t-1} \\ - (1-\delta)\left[I_{t-1} - (1/\alpha_1)(P_t - P_{t-1}) \right. \\ \left. - \delta\alpha_0/\alpha_1(1-\delta)\right]\end{aligned} \tag{15}$$

where the expression for I_t is obtained by eliminating P^e_{t+1} from (12), (13), and (14).[14] Equation (15) can readily be solved for P_t. Letting $\xi = \alpha_1/(1-\delta)$ and $k_1 = 1 + \xi(\gamma_1 - \beta_1)$, the result is

$$\begin{aligned}P_t = \frac{-\alpha_0\delta}{k_1(1-\delta)} + \frac{\xi\beta_2}{k_1} y_t - \frac{\xi\gamma_2}{k_1} w_t \\ + \frac{1}{k_1} P_{t-1} - \frac{\xi\delta}{k_1} I_{t-1}\end{aligned} \tag{16}$$

or more simply,

$$\begin{aligned}P_t = \pi_0 + \pi_1 y_t + \pi_2 w_t + \pi_3 P_{t-1} \\ + \pi_4 I_{t-1}\end{aligned} \tag{16'}$$

Intuitively acceptable signs for π are implied by the underlying parameters. These signs are reported in Table 1, where the present model is referred to as *INV*-I. It can be seen that this model is comparable to the one, based on the law of supply and demand, denoted *S&D*-I. The models are empirically distinguishable, however, since the latter implies $\pi_4 = 0$.

Let us now introduce alternative assumptions regarding price expectations. The adaptive expectations formula can be viewed as expressing P^e_{t+1} as a geometric distributed lag function of past actual values. It is natural to consider other distributed lag formulations. Of the numerous possibilities, two were chosen for study: (i) the second-order Pascal and (ii) a distribution with a geometric tail in which the first two weights are determined freely. With the relevant expressions used in place of (13), equations analogous to (16) can be derived. The implications are reported for *INV*-II and *INV*-III in Table 1.

---

[13] The latter condition assumes that the cost-of-storage function includes, as a negative component, a "convenience yield" that more than offsets explicit costs at low inventory levels; see Brennan (1959, pp. 238–39).

[14] We eliminate P^e_{t+1} before using the market-clearing condition since the immediate object is to derive a supply function.

We shall also consider two expectational assumptions that do not refer to distributed lag specifications. The first of these uses, in place of (13), the "extrapolative expectations" formula introduced by Lloyd Metzler. It yields model *INV*-IV. The implied signs depend upon a term that can be either positive or negative, but nevertheless the model has a differentiating implication: the coefficients attached to y_t and w_t have the *same* sign as I_{t-1}. Finally, we consider John Muth's "rational expectations" hypothesis. In this case there is no formula analogous to (13); instead we have the condition that P^e_{t+1} equal the mean of the probability distribution of P_t. This condition has implications only if assumptions are made regarding the nature of the stochastic process involved. If we adopt the "standard" assumption that each random shock is independent of all previous shocks, we can, following Robert Lucas, conjecture a solution of the form

$$P_t = \pi_0 + \pi_1 y_t + \pi_2 w_t + \pi_3 I_{t-1} \tag{17}$$

where $\pi_1, \pi_2 > 0$ and $\pi_3 < 0$. The distinguishing feature of the Muth-Lucas theory is the absence of P_{t-1} from the list of explanatory variables.

Thus we have five[15] plausible specifications concerning price expectation behavior. In conjunction with the supply-of-storage inventory theory, these have permitted derivation of five distinct estimating equations, no two of which have exactly the same implications regarding coefficient signs—again, see Table 1. Below we shall report estimates for each of these formulations.

III. Application to Lumber Industry

Having explained in general outline the models to be examined, we must now consider their application to a particular industry. The *U.S.* lumber manufacturing industry was chosen because of its unusually atomistic structure[16] and because lumber prices have recently been extremely volatile (and thus rich in information). Since our formulations are all expressed in real terms, the (quarterly) price movements to be explained are those of the wholesale price index for lumber, deflated by the index for all commodities.[17]

Implementation of the models developed above requires specification of demand and output equations more adequate than (6) and (7). The demand for lumber, to take the simpler case first, was taken to be a function of its price, real *GNP*, and the price of plywood, a close substitute. Thus with $\beta_1 < 0$ and $\beta_2, \beta_3 > 0$, we have

$$D_t = \beta_0 + \beta_1 P_t + \beta_2 GNP_t + \beta_3 PPLY_t \tag{18}$$

where the symbols have obvious meanings.

On the supply side, the counterpart of equation (7) must reflect the important determinants of (short-run) marginal cost. Obviously a wage rate (W) and the price of timber should be included. For the former, we took production worker average hourly earnings for Sawmills and Planing Mills (SIC 242). For the latter, the leading possibilities were provided by U.S. Forest Service quarterly data on average stumpage prices for various species of sawtimber. The particular series chosen was that for Southern Pine (Southern Region); the basis for this choice was the excellent explanation provided by southern pine stumpage prices (*PSP*) for

---

[15] Actually, eight expectational specifications were tried, each of the distributed lag formulations being applied to price changes, as well as levels. The empirical results were slightly inferior for each of the cases involving changes, however, so they have been omitted in order to shorten the discussion.

[16] In 1967 there were 10,271 establishments (owned by 10,016 firms) in SIC 2421, Sawmills and Planing Mills (see the *Census of Manufactures*). The largest eight companies accounted for 15 per cent of industry sales; the largest 50 for 31 percent.

[17] All other price series are similarly deflated. The figures come from the Bureau of Labor Statistics publication, *Wholesale Prices and Price Indexes*.

annual levels of material cost per unit of output in the lumber industry (SIC 2421).[18] In addition to variable input prices, the position of a competitive industry's short-run marginal cost curve depends upon the "quantity" of plant and equipment available. Since accounting data on fixed assets in lumber are sparse, available only annually, and conceptually inappropriate, an index of "capital" was constructed by a procedure, which I will furnish upon request, that reflects both technical progress and accumulation. With KC used to denote this index, the lumber output function can be expressed as

$$Q_t = \gamma_0 + \gamma_1 P_t + \gamma_2 W_t + \gamma_3 PSP_t + \gamma_4 KC_t \tag{19}$$

Here we should have $\gamma_2, \gamma_3 < 0$ and $\gamma_1, \gamma_4 > 0$.

With (18) and (19) replacing (6) and (7), operational counterparts of the various price equations are obtained by steps similar to those of Sections I and II. The counterpart of INV-I, to give one example, can be written

$$\begin{aligned} P_t = \pi_0 &+ \pi_1 W_t + \pi_2 PSP_t + \pi_3 KC_t \\ &+ \pi_4 GNP_t + \pi_5 PPLY_t \\ &+ \pi_6 P_{t-1} + \pi_7 I_{t-1} \end{aligned} \tag{20}$$

with implications[19] $\pi_1, \pi_2, \pi_4, \pi_5, \pi_6 > 0$ and $\pi_3, \pi_7 < 0$

IV. Empirical Results

Ordinary least squares estimates of the $S\&D$ and INV models, based on the 48 quarters of the period 1958–69, are reported in Tables 2 and 3, respectively.[20] Figures in parentheses, in these tables, are absolute values of t-ratios. Both adjusted ($\bar{R}^2$) and unadjusted (R^2) coefficients of determination are shown, together with unadjusted values ($R^2_{\Delta P}$) implied by the various equations for price *changes*. Values of James Durbin's statistic for testing for serial correlation in autoregressive models are denoted h. Estimates are not shown for constant terms or seasonal dummy variables, which were included in all cases.

Consider first the Supply and Demand formulations. The results for model $S\&D$-I, in which price responds to current excess demand, are quite favorable. Values for both R^2 and $R^2_{\Delta P}$ are high, and Durbin's h statistic is well below 1.645, the critical value for the usual test.[21] More importantly, all coefficients are of the hypothesized sign and each t-ratio exceeds 2.5 in absolute value—with all but one greater than 4.0. Model $S\&D$-II, which presumes a one-quarter lag in the price response, provides much less explanatory power. Indeed, the value of $R^2_{\Delta P}$ is so low for this specification that it will not be referred to again.

The two-quarter response model, $S\&D$-III, can also be eliminated since several of the coefficients are of the wrong sign and the h-statistic gives strong indication of first-order serial correlation in the disturbances. In the case of $S\&D$-IV, on the other hand, the results are again favorable. In this model, which presumes a geometric distributed lag response, P_{t-2} is added to the list of explanatory variables appearing in $S\&D$-I. The addition increases the equation's explanatory power, as is evidenced by the augmented value of $\bar{R}^2$ and the t-ratio of 3.6 for the coefficient attached to P_{t-2}. It is not obvious that

---

[18] The material cost/unit figures were calculated from census data. Regression of these figures for 1949–66 on annual PSP values, plus a dummy variable to take account of a 1954 change in the industry definition, produced on R^2 of 0.94. Douglas fir prices, another obvious possibility, provided much less explanatory power.

[19] Inventory figures were obtained from the Commerce Department's *Business Statistics*.

[20] Two-stage least squares estimates were also obtained, with W and PSP treated as jointly dependent variables. As the results were generally quite similar to those obtained via ordinary least squares, only the latter are here reported.

[21] The significance level here, and in all tests mentioned below, is 0.05.

TABLE 2—ESTIMATES FOR SUPPLY AND DEMAND MODELS

Variables	Coefficients and *t*-ratios[a]			
	S&D-I	*S&D*-II	*S&D*-III	*S&D*-IV
W	.2674		.3164	.2941
	(5.7)		(4.0)	(7.0)
W_{-1}		.5228	−.0518	
		(2.3)	(0.5)	
PSP	.4460		.5622	.4318
	(6.9)		(5.7)	(7.6)
PSP_{-1}		.5638	−.1225	
		(1.7)	(0.9)	
KC	−.2778		−.2224	−.1640
	(4.4)		(1.1)	(2.6)
KC_{-1}		−.4307	.1020	
		(1.8)	(0.5)	
GNP	.1484		−.0694	.0285
	(2.7)		(0.8)	(0.5)
GNP_{-1}		.1001	.2705	
		(0.5)	(2.6)	
$PPLY$	.4464		.4125	.4180
	(16.5)		(13.5)	(16.8)
$PPLY_{-1}$		.4028	.0615	
		(2.4)	(1.1)	
P_{-1}	.3663	.0906	.3298	.5175
	(9.0)	(0.3)	(3.0)	(9.3)
P_{-2}				−.1741
				(3.6)
		Statistics		
$\bar{R}^2$	.986	.842	.988	.989
R^2	.989	.872	.991	.992
$R^2_{\Delta P}$	.942	.347	.956	.957
h	1.19	[b]	3.73	1.79

[a] Not reported for constant term and seasonal dummies, which were included in each case.

[b] h-statistic not calculable because $n\hat{V}(b_1)>1$; $DW=1.55$.

this formulation is superior to *S&D*-I, nevertheless, as the *h*-statistic is slightly greater than its critical value and the *t*-ratio for the *GNP* variable is only 0.5. Consequently, we shall retain both *S&D*-I and *S&D*-IV in our list of contending specifications.

Next consider estimates, reported in Table 3, of the various inventory formulations. As models *INV*-I and *INV*-IV include the same list of explanatory variables, the first column applies to both. Its contents are sufficient to eliminate the latter version, which implies that the coefficient of I_{t-1} should be of the same sign as those attached to *W*, *PSP*, *GNP*, and *PPLY*. Model *INV*-V must also be ruled out, as it implies a zero value for the coefficient of P_{t-1}. Of the two remaining specifications, *INV*-II and *INV*-III, the latter appears to be slightly superior, as I_{t-2} fails to enter significantly in *INV*-II. Another reason for preferring *INV*-III is that its specification is the same as that of *S&D*-IV, except for the inclusion of I_{t-1}. Since *INV*-I and *S&D*-I are related in the same fashion, we have two possibilities for formal hypothesis testing.

TABLE 3—ESTIMATES FOR INVENTORY MODELS

Variables	Coefficients and t-ratios[a]			
	INV-I[b]	INV-II	INV-III	INV-V
W	.3028	.3396	.3307	.5354
	(6.7)	(8.5)	(8.6)	(7.9)
PSP	.3894	.3635	.3737	.7596
	(6.2)	(6.7)	(7.1)	(8.6)
KC	−.2145	−.0789	−.0975	−.2368
	(3.5)	(1.2)	(1.6)	(1.4)
GNP	.0227	−.1291	−.1018	−.1595
	(0.3)	(1.8)	(1.6)	(1.4)
$PPLY$	.4374	.4124	.4085	.4472
	(17.5)	(18.1)	(18.5)	(9.9)
P_{-1}	.3531	.5119	.5064	
	(9.3)	(10.3)	(10.3)	
P_{-2}		−.1933	−.1768	
		(4.0)	(4.1)	
I_{-1}	−.0211	−.0171	−.0216	−.0299
	(2.8)	(2.0)	(3.4)	(2.2)
I_{-2}		−.0076		
		(0.8)		
		Statistics		
$\bar{R}^2$	.988	.992	.992	.961
R^2	.991	.994	.994	.969
$R^2_{\Delta P}$	.952	.968	.967	.840
h	1.45	1.55	1.83	[c]

[a] Not reported for constant term and seasonal dummies, which were included in each case.

[b] Also model INV-IV.

[c] Not relevant; $DW=1.32$.

The law of supply and demand, it will be recalled, attempts to explain price changes solely on the basis of present and past levels of excess demand. It thus implies that I_{t-1} has no significant effect on price levels in either INV-I or INV-III. In contrast, our inventory theory predicts a negative effect in each case. A consistent test can, of course, be based on the t-ratios associated with I_{t-1}, which are asymptotically distributed as standard normal random variables. One-sided tests are appropriate, so the critical regions are $(-\infty, -1.645)$.

Consider, then, the results of Table 3. In both cases the test statistic lies well inside the critical region. The I_{t-1} t-ratios, that is, are negative and much larger than 1.645 in absolute value in both INV-I and INV-III. Thus we are able to reject the law of supply and demand hypotheses in favor of the more elaborate optimal inventory alternatives.

This result was sustained, furthermore, by additional evidence. Two of the variables used to represent nonprice determinants of supply and demand, namely W and $PPLY$, represent fairly obvious choices. In the other three cases, however, the particular choices, PSP, GNP, and KC, could reasonably be challenged. Accordingly, several experiments were carried out using alternative definitions. These suggested that our main results are not sensitive to changes in the definition of explanatory variables. In particular, such changes increased the absolute magnitude of the (negative) coefficients and *t*-ratios associated with I_{t-1}.[22]

Thus our evidence suggests that the price adjustment models based on supply-of-storage inventory theory, besides being theoretically more explicit and complete, are empirically superior to the inventory-free specifications provided by the law of supply and demand. Two qualifications should, however, be mentioned. First, it would clearly be wrong to base sweeping conclusions on results for only one industry, however ideally suited for the study. Second, it must be admitted that two of the supply-demand formulations were found to possess considerable explanatory power, and to be consistent with the evidence in all respects except for the implication that inventory levels "do not matter." Our results, while indicating superiority of the inventory models, do not contradict the view that the supply-demand law is a useful approximation, one that represents in a simplified way complex dynamic behavior.

---

[22] These results, as well as those mentioned in fn. 20, may be obtained from the author upon request.

REFERENCES

K. J. Arrow, "Towards a Theory of Price Adjustment," in *The Allocation of Economic Resources*, Stanford 1959.

R. Bentzel and B. Hansen, "On Recursiveness and Interdependency in Economic Models," *Rev. Econ. Stud.*, 1955, *22*, 153–68.

M. J. Brennan, "The Supply of Storage," *Amer. Econ. Rev.*, Mar. 1958, *48*, 50–72.

———, "A Model of Seasonal Inventories," *Econometrica*, Apr. 1959, *27*, 228–44.

T. J. Courchene, "An Analysis of the Price-Inventory Nexus with Empirical Application to the Canadian Manufacturing Sector," *Int. Econ. Rev.*, Oct. 1969, *10*, 315–36.

J. Durbin, "Testing for Serial Correlation in Least-Squares Regression when Some of the Regressors are Lagged Dependent Variables," *Econometrica*, May 1970, *38*, 410–21.

O. Eckstein and G. Fromm, "The Price Equation," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1159–83.

R. C. Fair, *A Short-Run Forecasting Model of the United States Economy*, Lexington 1971.

D. F. Gordon and A. Hynes, "On the Theory of Price Dynamics," in E. S. Phelps, ed., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

A. C. Harberger, *The Demand for Durable Goods*, Chicago 1960.

G. A. Hay, "Production, Price, and Inventory Theory," *Amer. Econ. Rev.*, Sept. 1970, *60*, 531–45.

R. E. Lucas, "Econometric Testing of the Natural Rate Hypothesis," in O. Eckstein, ed., *The Econometrics of Price Determination*, Washington 1972.

B. T. McCallum, "The Effect of Demand on Prices in British Manufacturing: Another View," *Rev. Econ. Stud.*, Jan. 1970, *37*, 147–56.

———, "Inventory Holdings, Rational Expectations, and The Law of Supply and Demand," *J. Polit. Econ.*, Mar./Apr. 1972, *80*, 386–93.

L. A. Metzler, "The Nature and Stability of Inventory Cycles," *Rev. Econ. Statist.*, Aug. 1941 *22*, 113–29.

E. S. Mills, *Price, Output and Inventory Policy*, New York 1962.

J. F. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

R. R. Neild, *Pricing and Employment in the Trade Cycle*, Cambridge 1963.

M. Nerlove, *Distributed Lags and Demand Analysis for Agricultural and Other Commodities*, Washington 1958.

F. Rushdy and P. J. Lund, "The Effect of Demand on Prices in British Manufacturing Industry," *Rev. Econ. Stud.*, Oct. 1967, *34*, 361–73.

C. L. Schultz and J. L. Tryon, "Prices and Wages," in J. S. Duesenberry, et al., eds., *The Brookings Quarterly Econometric Model of the United States*, Amsterdam 1965.

L. G. Telser, "Futures Trading and the Storage of Cotton and Wheat," *J. Polit. Econ.*, June 1958, *66*, 233–55.

H. Working, "The Theory of Price of Storage," *Amer. Econ. Rev.* Dec. 1949, *39*, 1254–62.

W. J. Yordan, "Industrial Concentration and Price Flexibility in Inflation: Price Response Rates in Fourteen Industries, 1947–1958," *Rev. Econ. Statist.*, Aug. 1961, *43*, 287–94.

U.S. Bureau of the Census, *Census of Manufactures*, Washington, various issues.

U.S. Bureau of Labor Statistics, *Wholesale Prices and Price Indexes*, Washington, various issues.

———, *Employment and Earning Statistics for the United States*, Washington, various issues.

U.S. Department of Agriculture, Forest Service, *The Demand and Price Situation for Forest Products*, Washington, various issues.

U.S. Office of Business Economics, *Business Statistics*, Washington, various issues.

An Economic Model of Production and Investment for Petroleum Reservoirs

By Robert G. Kuller and Ronald G. Cummings*

Economists have long been concerned with problems associated with petroleum production. However, efforts to develop an analytical framework for the study of such problems have been impeded by the morass of technological and socioeconomic problems peculiar to this decision environment. Analytical models suggested in earlier works do not consider certain processes which are basic to the technology of petroleum production. For example, Paul Davidson, James McKie and Stephen McDonald, and Vernon Smith, fail to provide an analytical model for petroleum production which includes interaction among (a) current production rates, (b) the role of investment (both in terms of pressure maintenance, and present and/or future use of artificial lifting methods), and perhaps most important, (c) the possible dependence of recoverable stocks on the time-path of production. Further, earlier works, with the exception of those by Smith and Kuller, rely on linear production paths, and only the latter attempts to develop explicitly both the economic and physical nature of the user costs involved in petroleum production.

The purpose of this paper is to present an economic model for petroleum reservoir management which includes the above described characteristics of petroleum production. The model, essentially a specific application of the general resource model developed by Oscar Burt and Cummings, represents our attempt to characterize production relations in a manner consistent with current procedures used by petroleum engineers for predicting reservoir performance. The model is designed with a combination water-gas drive reservoir in mind, yet is adaptable for application to other types of reservoirs.

In Section I below, the arguments begin with a physical description of a petroleum reservoir, in which an attempt is made to find a middle ground for exposition that is sufficiently technical to satisfy readers familiar with petroleum technology and yet is comprehensible to other readers.

In Section II, the management model is presented; the mathematical development for this model is included in the Appendix. An economic interpretation of the resulting decision rules for production and investment is given in Section III, and an extended discussion of the implications of such rules for policy, as well as a comparison of results from earlier works, is the topic of Section IV. Concluding remarks are given in Section V, along with suggestions concerning the possible extension of this management model for the purpose of determining optimal expenditures for exploration.

I. Petroleum Reservoir Mechanics

A petroleum reservoir is a body of porous and permeable rock saturated with a solution of oil, gas, and ordinarily, water. An untapped reservoir exists under very high pressure, imposed on the reservoir by, for example, confining layers of impermeable rock, an aquifer which surrounds the reservoir, a gas cap or pocket, or some combination of the preceding.

---

* Assistant professor of economics, University of Wyoming, and professor of resource economics, University of Rhode Island, respectively. We wish to express our appreciation to Resources for the Future, Inc., The Division of Business and Economic Research, University of Wyoming, and the Rhode Island Agricultural Experiment Station (Contribution #1510) for their generous support of this work.

When the reservoir is tapped by a well, an immediate pressure differential results. The pressure in the well-hole is considerably less than the pressure in the reservoir. As a result, the tightly compressed molecules of oil and gas expand and move in the direction of the well. Similarly, it is the pressure differential between the bottom and top of the well which, if great enough, gives rise to the continued "natural" flow of the oil-gas solution to the surface. Production of oil and gas by means of this phenomenon is referred to as production from the reservoir's "natural drive."

The amount of pressure which exists in the reservoir, and particularly, the time-path of pressure depletion, are crucial in terms of two important dimensions of petroleum production, viz., the production rate and the amount of oil recoverable from the reservoir. In terms of production, the relevance of sufficient pressure to provide production from natural drive is immediately obvious. Such production is virtually costless in terms of direct variable costs (that such production is not costless, in terms of *user* costs, is developed in the following two sections). Even when natural drive forces alone are no longer capable of sustaining production, the extent of artificial lift production (for example, pumping) is affected by the rate at which oil moves through the reservoir into the well by natural forces. This rate depends upon the amount of dissolved gas in the solution, which in turn is dependent upon the previous time-path of pressure.

The recoverable portion of the total volume of oil in the reservoir depends upon a number of factors. A primary determinant is the source of the reservoir's natural drive: dissolved gas drive (production derives solely from the expansion of oil and gas), gas cap drive (the expansion of a free gas "pocket" forces the oil toward the well), water drive (water displaces the oil and maintains pressure), or combination drive which results from a combination of the preceding forces. The portion of the initial oil stock that is recoverable by natural drive will vary from 15 to 30 percent in dissolved gas reservoirs to as high as 80 percent in gas cap, water and combination drive reservoirs (see N. Clark, Part 1, p. 160).[1]

Of particular importance in this paper is the fact that the recoverable stock may be *rate dependent*; i.e., the recoverable stock may depend upon the time-path of production. Following Benjamin Craft and Murray Hawkins, ". . . many reservoirs are clearly *rate sensitive* and there is a *maximum efficient rate* (MER) above which there will be a significant reduction in the practical ultimate oil recovery" (p. 197). Such dependence is particularly relevant for water drive and combination drive reservoirs.

Consider, for example, the mechanics of a water drive reservoir. As pressure in the reservoir is reduced by production, water displaces the oil ahead of it, essentially maintaining an interface, or juxtaposition, with the body of oil (i.e., the body of water continues to "push against" the body of oil). Thus, pressure is maintained, and ". . . provided that water moves through all segments of the reservoir sand and all areas over which the reservoir occurs" (Clark, Part 1, p. 160), only small amounts of oil are lost in isolated pockets and in mixture with water, so that a large portion of oil is gleaned from the saturated reservoir rock.

---

[1] The relatively low percentage recovery by dissolved gas drive follows from the fact that production by this method implies the simultaneous production of the reservoir's principle source of pressure—not necessarily the case for production by the other mechanisms.

The rate at which water will displace the oil, and thereby maintain pressure, is in turn determined by a number of factors, for example, the relative size of the aquifer and amount of its pressure, and the permeability and porosity of the reservoir rock. If oil is extracted at a greater rate than that at which water displaces the oil, reservoir pressure is reduced. Also, "since water moves more freely than oil through the tiny pores and interstices of oil-bearing rock, too rapid production of oil . . . may allow water to surround and bypass oil in place, which usually means permanent loss" (McKie and McDonald, p. 102).[2] For example, suppose 5 million barrels of oil are to be extracted over a 5-year period. A uniform production rate of 1 million barrels per year might have a considerably different effect on the ultimate recoverable stock than a sequence of rates of 4 million barrels the first year and 250,000 barrels each of the following 4 years. Thus, the essence of rate dependence is that total recovery, as well as annual production rates from natural drive, may depend not only on cumulative production, but also on the rate at which such production has taken, and will take place (see Clark, Part 1, p. 160; Robert Gordon, p. 278, n. 8; McKie and McDonald, pp. 102–03; S. J. Pirson, pp. 708–10; and Smith, p. 414, n. 6).

From the above discussion it is obvious that pressure is a key element in petroleum production. While the discussion thus far has been focused on "natural" pressure in the reservoir, it should be noted that in many cases pressure may be maintained or increased artificially through investment activities. Water, gas, steam, or other fluids may be pumped into the reservoir (see Pirson, pp. 539–42). Thus, the recoverable stock, as well as production rates, generally will depend strongly on investment. Such dependence is not only on cumulative investment (i.e., the capital stock), but also on the time-path of investment (i.e., there is rate dependence with reference to investment). The rate dependence of recoverable stocks on investment follows directly from the above argument concerning production rates. Thus, the impact of rapid production rates on total recovery may be offset to some extent by investments for pressure maintenance.

The evaluation and prediction of reservoir performance requires considerable data concerning reservoir pressure, geological properties of the reservoir rock, fluid properties, and production history (for example, see A. Baca, p. 1; Clark, Part 1, p. 157; and Pirson, p. 533). These data are then used by engineers to develop "reservoir performance curves" similar to those depicted in Figure 1. Given an annual rate of extraction, for simplicity some constant value u', an estimate is made of the future time-path of pressure maintained in the reservoir. In this manner the "production life," and therefore the recoverable stock from the reservoir with reference to natural drive is obtained. In many cases, a faster annual extraction rate, u'' in Figure 1, may result in a leftward movement in the cumulative production curve, reflecting as discussed above, the "rate dependence" of recoverable stocks. If ultimate recovery is rate dependent the faster rate u'' may not result in lower recovery given some annual investment, v', for pressure maintenance activities.

Thus, for some planning horizon of T periods[3] t = 1, 2, . . . , T, let:

---

[2] It is of interest to note that water drive *gas* reservoirs may behave in an opposite fashion, i.e., too slow production of gas may result in the loss of a significant quantity of the initial gas stock (see R. G. Agarwal, R. Al-Hussainy and H. J. Ramey).

[3] Comments on the choice of T are given in Section V below.

u_t = the total volume of petroleum extracted from the reservoir during period t

$U_t = (u_1, u_2, \ldots, u_t)$

v_t = gross investment for all capital components during period t

$V_t = (v_1, v_2, \ldots, v_t)$

X = the total recoverable quantity of petroleum in the reservoir

K_t = capital stocks at the beginning of t

F_t = an upper (physical) bound on the capacity to produce petroleum from the reservoir during period t

C_t = the generalized cost function for production in period t

The following general relations may then be postulated as conceptually determinable from the reservoir performance curve methodology used by petroleum engineers:[4]

$$(1) \qquad \sum_{\tau=1}^{T} u_\tau \leqq X(U_T, V_T),$$

$$\frac{\partial X}{\partial u_\tau} \leqq 0, \qquad \frac{\partial X}{\partial v_\tau} \geqq 0$$

$$(2) \qquad u_t \leqq F_t(U_t, V_t, K_t), \qquad \frac{\partial F_t}{\partial u_t} \leqq 0,$$

$$\frac{\partial F_t}{\partial v_t} \geqq 0, \qquad \frac{\partial F_t}{\partial K_t} \geqq 0$$

$$(3) \qquad C_t(U_t, V_t, K_t), \qquad \frac{\partial C_t}{\partial u_\tau} \geqq 0,$$

$$\frac{\partial C_t}{\partial v_t} \geqq 0, \qquad \underset{(\tau \neq t)}{\frac{\partial C_t}{\partial v_\tau}} \leqq 0, \qquad \frac{\partial C_t}{\partial K_t} \leqq 0$$

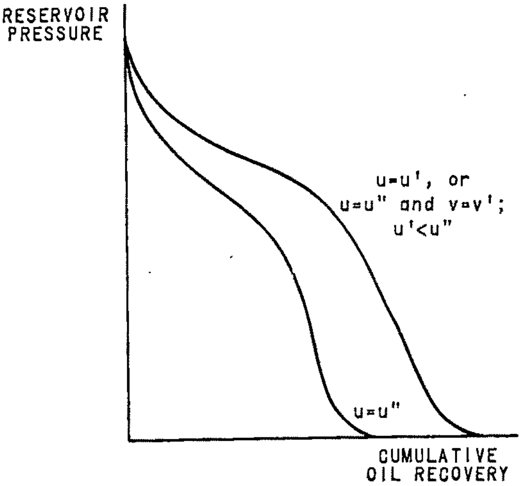


FIGURE 1

Equation (1) states that total petroleum production over the T period horizon may not exceed the recoverable stock X, where X is determined by the time-path of production and investment over the entire period $1 \leqq t \leqq T$. Since the dissolved gas content of the reservoir solution, as described above, affects the natural forces driving petroleum toward the well (for production by artificial lift), X is viewed here as recoverable stock from natural drive as well as from artificial lift.

Equation (2) imposes an upper bound, F_t, on annual production from the reservoir. This value of F_t is determined by reservoir pressure at t which is in turn determined by the previous and current values of u and v, and by capital stocks (for example, size and number of wells, and artificial lift equipment, in the case of artificial lift).

Equation (3) is a generalized cost function. Variable production costs in t may be (essentially) zero,[5] if sufficient pressure exists for production via natural drive, thus the dependence of C_t on the previous paths of u and v (reference to V_t here is to $v_1, \ldots, v_{t-1}$, where v_t is viewed as a current investment charge, i.e., $\partial C_t/\partial v_t \geqq 0$). If u_t and v_t are such that pumping is required in t, u_t results in current direct costs; i.e., $\partial C_t/\partial u_t > 0$. Capital stocks, reflecting the efficiency of the production system, are assumed to decrease unit costs.

---

[4] We gratefully acknowledge comments on an earlier draft of this paper by Oscar Burt. He suggests an alternative statement of the relations below which might be most useful in terms of empirical applications with dynamic programming, viz., the use of X_{1t}, reservoir pressure at t, and X_{2t}, maximum recoverable stock at t, as state variables, along with K_t, in a recursive programming format. We have experimented with this system and find that for analytical purposes the resulting decisions rules are almost identical. However, Burt's method may be preferred for empirical applications.

[5] Reference here is to the individual firm. This assumption may or may not reconcile James Griffin's recent arguments, but is of no concern at this point. The arguments, as shown below, are unaffected by rising marginal costs.

Having specified the above relations, (1) through (3), a production and investment model for the reservoir may now be suggested. The notation used above is expanded as follows:

u_{jt} = the volume of petroleum extracted by firm j, $j=1, \ldots, n$ during period t

U_t = annual production rates by *all* firms during all periods, 1, . . . , t; i.e., $(u_{11}, u_{21}, \ldots, u_{n1}, \ldots, u_{1,t-1}, \ldots, u_{n,t-1}, u_{1t}, \ldots, u_{nt})$

v_{jkt} = gross investment by firm j in capital component k, $k=1, \ldots, q$, during period t

$v_{jt} = (v_{j1t}, \ldots, v_{jqt})$

V_t = gross investment for all capital components by all firms during the periods 1, . . . , t

K_{jkt} = firm j's stock of capital component k at the beginning of period t

$K_{jt} = (K_{j1t}, \ldots, K_{jqt})$

D_{jkt} = net depreciation of firm j's stock of capital component k during period t

F_{jt} = an upper (physical) bound on firm j's capacity to produce petroleum during period t

C_{jt} = firm j's cost function during period t

β_t = a discount factor, $(1+r)^{-t}$, where r is the appropriate discount rate

p_t = unit price of petroleum during period t

II. A Management Model for the Reservoir

The existence of n firms exploiting a given petroleum reservoir is postulated. Production from the reservoir is assumed to be a sufficiently small part of total (national) petroleum production such that the n firms may be viewed as price takers. For a T period horizon, t=1, . . . , T, it is assumed that all prices are known.

Each firm j is assumed to maximize the value of a stream of discounted profits from the sale of petroleum over the planning horizon, which is equivalent to maximizing the following relation:

$$\sum_{t=1}^{T} [p_t u_{jt} - C_{jt}(U_t, V_t, K_{jt})]\beta_t \tag{4}$$

In the above, $p_t u_{jt}$ represents gross revenues to the firm; C_{jt} is the firm's cost function, the nature of which was detailed above in Section I.

The following relations and restrictions are assumed to apply to the production process

$$K_{jk,t+1} = K_{jkt} - D_{jkt}(u_{jt}, v_{jkt}, K_{jkt}) \tag{5}$$

and

$$u_{jt} \leqq F_{jt}(U_t, V_t, K_{jt}), \quad \text{for all } j, k, \text{ and t} \tag{6}$$

Equation (5) states that the stock of capital component k at the end of period t (i.e., the beginning of period t+1) equals the initial stock minus *net* depreciation. Net depreciation is assumed to decrease with gross investment and to increase with u and K. In most applications, either u, implying a consumptive-use notion of depreciation, or K, implying a capital-decay notion of depreciation, will be used. In either case, the following conditions hold:

$$\frac{\partial D_{jkt}}{\partial v_{jkt}} \leqq 0, \quad \frac{\partial D_{jkt}}{\partial u_{jt}} \geqq 0, \quad \frac{\partial D_{jkt}}{\partial K_{jkt}} \geqq 0 \tag{7}$$

In (6) the generalized upper bound on current production rates (in t) discussed above in equation (2) is imposed on the individual firm j. It should be noted that this bound relevant for firm j depends upon past and current production and investment rates from all firms producing from the reservoir (inasmuch as reservoir pressure is determined by total production from the reservoir, as well as from aggregate investment activities for pressure maintenance). Thus,

$$\frac{\partial F_{j\mathrm{t}}}{\partial u_{i\tau}} \leqq 0 \tag{8.1}$$

$$\frac{\partial F_{j\mathrm{t}}}{\partial v_{i\tau}} \geqq 0 \tag{8.2}$$

$$\frac{\partial F_{j\mathrm{t}}}{\partial K_{j\mathrm{t}}} \geqq 0, \tag{8.3}$$

for all i, j, k, and t; $\tau = 1, \ldots, \mathrm{t}$

Finally, for the reservoir as a whole, the following form of equation (1) discussed above is required in order to determine the recoverable stock:

$$\sum_{\tau=1}^{\mathrm{T}} \sum_{j=1}^{n} u_{j\tau} \leqq X(U_{\mathrm{T}}, V_{\mathrm{T}}) \tag{9}$$

Concern in this paper is with centralized control or management of the reservoir.[6] Thus, the relevant optimization problem is expressed by the following:

Maximize:

$$\sum_{\mathrm{t}=1}^{\mathrm{T}} \sum_{j=1}^{n} \{p_{\mathrm{t}} u_{j\mathrm{t}} - C_{j\mathrm{t}}(U_{\mathrm{t}}, V_{\mathrm{t}}, K_{j\mathrm{t}})\} \beta_{\mathrm{t}} \tag{10}$$

subject to the conditions

$$K_{jk,\mathrm{t}+1} = K_{jk\mathrm{t}} - D_{jk\mathrm{t}}(u_{j\mathrm{t}}, v_{jk\mathrm{t}}, K_{jk\mathrm{t}}) \tag{11}$$

$$u_{j\mathrm{t}} \leqq F_{j\mathrm{t}}(U_{\mathrm{t}}, V_{\mathrm{t}}, K_{j\mathrm{t}}) \tag{12}$$

$$\sum_{\tau=1}^{\mathrm{T}} \sum_{j=1}^{n} u_{j\tau} \leqq X(U_{\mathrm{T}}, V_{\mathrm{T}}) \tag{13}$$

$$u_{j\mathrm{t}} \geqq 0, \quad v_{jk\mathrm{t}} \geqq 0, \quad \text{for all } j, k, \text{ and t} \tag{14}$$

Necessary conditions for a maximum in the system (10) through (14) are developed in the Appendix. Discussion here continues with an economic interpretation of the decision rules implied by the conditions (A2) and (A3) in the Appendix.

III. Optimal Paths for Production and Investment

Rearrangement of terms in Appendix equations (A2) and (A3) yields the following relations which characterize optimal paths for production and investment for the reservoir during all periods t, $1 \leqq \mathrm{t} \leqq \mathrm{T}$, for which $u_{j\mathrm{t}}$ and $v_{j\mathrm{t}}$ are positive.

$$\begin{aligned} (15) \quad \left(p_{\mathrm{t}} - \frac{\partial C_{j\mathrm{t}}}{\partial u_{j\mathrm{t}}}\right)\beta_{\mathrm{t}} &= \lambda\beta_{\mathrm{T}}\left(1 - \frac{\partial X}{\partial u_{j\mathrm{t}}}\right) + \Psi_{j\mathrm{t}}\beta_{\mathrm{t}} \\ &- \sum_{\tau=\mathrm{t}}^{\mathrm{T}} \sum_{i=1}^{n} \Psi_{i\tau} \frac{\partial F_{i\tau}}{\partial u_{j\mathrm{t}}} \beta_{\tau} \\ &+ \sum_{k=1}^{q} \Delta_{jk,\mathrm{t}+1}\beta_{\mathrm{t}+1} \frac{\partial D_{jk\mathrm{t}}}{\partial u_{j\mathrm{t}}} \\ &+ \sum_{\tau=\mathrm{t}+1}^{\mathrm{T}} \frac{\partial C_{j\tau}}{\partial u_{j\mathrm{t}}} \beta_{\tau} + \sum_{\tau=\mathrm{t}}^{\mathrm{T}} \sum_{\substack{i=1 \\ i\neq j}}^{n} \frac{\partial C_{i\tau}}{\partial u_{j\mathrm{t}}} \beta_{\tau} \end{aligned}$$

$i, j = 1, \ldots, n; 1 \leqq \mathrm{t} \leqq \mathrm{T}$

$$\begin{aligned} (16) \quad \frac{\partial C_{j\mathrm{t}}}{\partial v_{jk\mathrm{t}}} \beta_{\mathrm{t}} &= -\Delta_{jk,\mathrm{t}+1}\beta_{\mathrm{t}+1} \frac{\partial D_{jk\mathrm{t}}}{\partial v_{jk\mathrm{t}}} + \lambda\beta_{\mathrm{T}} \frac{\partial X}{\partial v_{jk\mathrm{t}}} \\ &+ \sum_{\tau=\mathrm{t}}^{\mathrm{T}} \sum_{i=1}^{n} \Psi_{i\tau}\beta_{\tau} \frac{\partial F_{i\tau}}{\partial v_{jk\mathrm{t}}} \end{aligned}$$

---

[6] As suggested to us by the referee, the objective function of this paper may be limited in terms of social control over reservoir production. More general treatments are implied in Burt and Cummings, and Smith.

$$-\sum_{\substack{i=1\\ i\neq j}}^{n}\frac{\partial C_{it}}{\partial v_{jkt}}\beta_t-\sum_{\tau=t+1}^{T}\sum_{i=1}^{n}\frac{\partial C_{i\tau}}{\partial v_{jkt}}\beta_\tau$$

$$i, j = 1, \ldots, n; k = 1, \ldots, q; 1 \leqq t \leqq T$$

Characteristics of Optimum Production Rates

At any t, $1 \leqq t \leqq T$, any firm j produces at the rate described by (15), assuming that $u_{jt} > 0$; conditions for a zero production rate are discussed below. The left-hand side of (15) is, of course, the present value of marginal net income to j during period t. On the right-hand side of (15) are six terms which may be interpreted as the "user costs" of an increment in production during t by firm j. User costs are defined by Anthony Scott as "... the present value of future profit foregone by a decision to produce a unit of output today" (1967, p. 34). In the present context, "future profit" corresponds to future net incomes to the reservoir. Given the major role and peculiarities of user costs in petroleum production, it will be useful to develop in some detail the elements of such costs.

Stock User Costs

The term $\lambda\beta_T$ may be interpreted as the marginal scarcity value of petroleum to the reservoir as a whole for the entire period $t = 1, \ldots, T$. From Appendix equation (A7), λ may be nonzero only in the case where equality holds in the relation (13); i.e., the endogenously determined recoverable stock X is exhausted at or before T. The coefficient λ then measures the increase in net incomes from the the reservoir associated with an incremental change in the endogenously determined stock. This value is weighted by the term $(1-\partial X/\partial u_{jt})$ which is no less than one, since by equation (1) $\partial X/\partial u \leqq 0$ which reflects the impact on the total recoverable reservoir stock (through for example, pressure reduction) of an increment in production by firm j in t. Thus, the term $\lambda\beta_T(1-\partial X/\partial u_{jt})$ may be thought of as j's "stock user cost" during t.

Boundary User Costs

The Lagrangian multiplier Ψ_{it}, $i = 1, \ldots, n$, is associated with the above restriction set (12) which imposes an upper bound F_{it} on each firm i's production rate during any period $t = 1, \ldots, T$. Then for any period τ, $t \leqq \tau \leqq T$, during which $F_{i\tau}$ is effective—equality holds in (12)—for any firm i, $\Psi_{i\tau}$ measures the increase in net incomes which would result from an incremental relaxation of the restriction, and $\partial F_{i\tau}/\partial u_{jt}$ measures the "relaxation" obtained from an incremental change (lowering) of the firm j's production in t. Thus, the second and third terms on the right-hand side of (15) are interpreted as "boundary user costs."

User Costs of Capital Consumption

The multiplier $\Delta_{jk,t+1}$, for which an expanded expression is developed in Appendix equation (A11), measures the marginal productivity of capital-type k used by firm j in all future periods $t+1$, $t+2, \ldots, T$. The term $\partial D_{jkt}/\partial u_{jt}$ measures the capital consumed in t (and therefore not available, *ceteris paribus*, in all future periods) by an increment in production by j in t. Thus, $\Delta_{jk,t+1}\beta_{t+1}\partial D_{jkt}/\partial u_{jt}$ measures the user cost of capital consumption by firm j.

Production User Costs

As discussed in Section I above, given a set of investment rates, current production may affect natural drive pressures as well as the characteristics (for example, viscosity) of the oil-gas solution for artificial lift purposes in current and/or future years. The resulting impacts of current production (for example, movement from natural drive to pump production, lower pump efficiencies from high viscosity solutions) on j's (as well as all other firms') future variable costs are the measures given by the last two terms on the right-hand side of (15), and thus, their interpretation as production user costs. These user costs reflect the stock value of oil and gas to the firm, contributing to output as natural factors of production (see for example, James Quirk and Smith, p. 9, for a discussion of the effects of resource stocks on output).

In summary, the rate characterized by (15) is that which follows the rule: produce at a rate where marginal net income to firm j equals the user cost associated with firm j's production. The user cost reflects not only the impact of u_{jt} on future costs internal to j, but also all external effects which result from j's production, viz., effects on the aggregate recoverable stock (λ), effects on other firms' maximum rates of production (Ψ), and impacts on all firms' direct production costs ($\partial C_i/\partial u_j$, $i=1, \ldots, n$). Of course, at any t during which inequality holds in (15), (i.e., marginal net income for j is less than j's user costs for all possible values of u_{jt}), $u_{jt}=0$ by Appendix equation (A2′), (i.e., firm j does not produce during t).

A further development of the ramifications of (15) for reservoir management requires a discussion of the characteristics of optimal investment rates, given the "offsetting" role in terms of pressure maintenance which investment activities vis-à-vis current production may play. This development is given in Section IV.

Characteristics of Optimum Investment Rates

Equation (16) dictates the optimal level of firm j's investment in capital-type k during any t, $1 \leqq t \leqq T$, for which such investment is positive. On the left-hand side of (16) is the present value of the marginal costs of such investment to firm j. On the right-hand side of (16) are aggregate benefits to the reservoir as a whole associated with such investment, and these benefits include the following elements. The first term is $\Delta_{jk,t+1}\beta_{t+1}$ weighted by $\partial D_{jkt}/\partial v_{jkt}$, the latter of which may be viewed as equal to -1. The term Δ_{jkt} is expanded in Appendix equation (A11) and repeated here, adjusted to t+1.

$$(17)\quad \Delta_{jk,t+1}\beta_{t+1} = \sum_{\tau=t+1}^{T}\left(-\frac{\partial C_{j\tau}}{\partial K_{jk\tau}} + \Psi_{j\tau}\frac{\partial F_{j\tau}}{\partial K_{jk\tau}}\right)\beta_\tau \prod_{s=t+1}^{\tau-1}\left(1-\frac{\partial D_{jks}}{\partial K_{jks}}\right) + \prod_{s=t+1}^{T}\left(1-\frac{\partial D_{jks}}{\partial K_{jks}}\right)\Delta_{jk,T+1}\beta_{T+1}$$

In (17), the $\partial C_j/\partial K_{jk}$, nonpositive by (3), are direct effects on j's extraction costs of larger capital stocks; $\Psi_j(\partial F_j/\partial K_{jkt})$ is the imputed value of capital by way of relaxing the bound F (see equation (2)). Then the sum is the present value of the stream of benefits to j of an increment in v_{jk} in t, discounted by a capital survival factor $(1-\partial D/\partial K)$. The last term in (17) may be viewed as the marginal value of current investment in terms of terminal capital stocks, which in most applications will be zero. Equation (17) may thus be interpreted as the present value of the future stream of direct benefits (the marginal productivity) from an increment in investment during t.

The second term in (16) measures external impacts of an increment in j's investment in t. Thus $\lambda\beta_T$ is the marginal scarcity value of recoverable stocks, and $\partial X/\partial v_{jkt}$ is the impact on such stocks of an increment in j's investment in capital-type k in t. The third term measures for all firms, the above impact (pressure maintenance) vis-à-vis the upper bound F. The fourth and fifth terms measure, for firm j and all other firms, the marginal impact of investment in capital-type k with reference to pressure maintenance and therefore, operating (variable) costs, for all periods $\tau \geqq t$.

The rule (16) then has the interpretation: equate the marginal cost of investment (to firm j for capital-type k) with the marginal present value of the reservoir-wide benefits associated with such investment, where the latter includes not only direct impacts on the marginal productivity of j's capital and j's future variable and boundary costs, but also external impacts on other firms' variable and boundary costs as well as on the recoverable stock.

Of course, for many types of capital (elements of $k=1, \ldots, q$), several elements in (16) will not apply. For investment in pressure maintenance activities, all terms will probably be included. For later investments in, for example, pumping equipment, it may be that only the first term will appear.

With the above characterizations for optimum production and investment rates, attention may now be turned to a brief discussion of the ramifications of such rates for reservoir management policies.

IV. Policy Ramifications

Consider the decision environment for petroleum extraction during the early "rule of capture" period in the United States (see McKie and McDonald, p. 98). Under such highly competitive production conditions (for a single reservoir), each firm is concerned only with the cost-benefit items directly affecting it. Efforts by a firm j to restrain production and/or to incur investment costs for pressure maintenance, for the purpose of increasing recoverable stocks and lowering future production costs (by maintaining natural drive forces), may be negated by the lack of such restraint on the part of other firms resulting in losses accruing to j. These potential losses imply, among other things, a very high discount rate, which in turn implies relatively low values for those terms in (15) and (16) with values depending on future flows. To clarify the point under discussion, suppose that firm j ignores all externalities. The relevant optimization problem is given in Appendix equations (A12) and (A13). Conditions which correspond to those given in (15) and (16) are as follows:

$$(18)\quad \left(p_t-\frac{\partial C_{jt}}{\partial u_{jt}}\right)-\sum_{k=1}^{q}\Delta_{jk,t+1}\frac{\partial D_{jkt}}{\partial u_{jt}}\beta_{t+1}=\Psi_{jt}$$

and

$$(19)\quad -\frac{\partial C_{jt}}{\partial v_{jkt}}=\Delta_{jk,t+1}\frac{\partial D_{jkt}}{\partial v_{jkt}}\beta_{t+1} .$$

for some $j = 1, \ldots, n;\ 1 \leqq t \leqq T$

Firm j is thus seen to be producing "rationally" at that rate where $u=F$ (see equation (2)), in which case marginal profit (which is essentially the unit price of petroleum under natural drive conditions) less capital consumption costs equals the marginal imputed cost of the upper bound F, which in this case depends solely on $\bar{K}_{jt}$. This bound may be relaxed by investment, for example, new wells (which is obviously limited by physical restrictions not included in this model). Production is thus at the maximum rate (F) allowed by producing facilities (wells), and as long as the marginal costs of expanding such facilities are less than the marginal future value of larger future production facilities, such facilities are expanded to their physical limit.

The impact of such production patterns in terms of losses, for example, in recoverable stocks, is self-evident. Central management of the reservoir must increase the present value of the income stream attributable to reservoir production. Such control may take the form of production "rules" based on the rates implied by (15) and (16), or the form of unit-taxes which capture all user cost items on the right-hand sides of (15) and (16).

In terms of taxes, it has been argued that depletion allowances have the effect of increasing not only production, but also investment rates (see for example, Agria, pp. 92–100; Harberger, p. 444; McDonald (1967), p. 280; and Steiner, p. 951, n. 6). Setting aside questions regarding the desirability of such increases, the effects may be demonstrated quite easily by (15) and (16). Let *UC* denote all user cost elements on the right-hand side of (15), and let *MP* denote all future productivity effects of investment v_{jkt} on the right-hand side of (16), excluding in both cases, the terms involving $\lambda\beta_T$. Solving (16) for $\lambda\beta_T$ and substituting into (15), results in the following expression:

$$\frac{\left(p_t - \dfrac{\partial C_{jt}}{\partial u_{jt}}\right)\beta_t - UC}{\left(1 - \dfrac{\partial X}{\partial u_{jt}}\right)} = \frac{\dfrac{\partial C_{jt}}{\partial v_{jkt}}\beta_t - MP}{\dfrac{\partial X}{\partial v_{jkt}}} \tag{15'}$$

If $(p_t - \partial C_{jt}/\partial u_{jt})$ is viewed as "after-tax" marginal profits, equality holds in (15′) for some value $\bar{u}_{jt}$ and $\bar{v}_{jkt}$. A depletion allowance increases the value of marginal profits at $\bar{u}_{jt}$, in which case inequality results in (15′). By Appendix equation (A2), this inequality implies an increase in u_{jt} to some value $\bar{\bar{u}}_{jt} \geqq \bar{u}_{jt}$. As u_{jt} rises, elements in the *MP* term may rise, in which case inequality would hold in (16), and by Appendix equation (A3), equality would be restored only by an increase in v_{jkt} to some value $\bar{\bar{v}}_{jkt} \geqq \bar{v}_{jkt}$. Thus, depletion allowances may indeed result in increases in production and investment rates (for some capital items *k*). This analysis suggests that tax provisions which tend to increase marginal after-tax profits (for example, the unit of production method of depreciating tangible exploration and development costs) may lead to a similar result (see McDonald (1967), p. 279). However, policies which tend to reduce marginal after-tax profits (for example, the imposition of a severance tax) may be expected to reduce production and investment rates (see Agria, pp. 101–04).

The results from this work may be compared with those from earlier works as follows. The present model extends the work of McKie and McDonald by (a) allowing many firms to produce from a reservoir; (b) allowing optimal production and investment rates to vary over the economic life of the reservoir; (c) allowing investment in artificial lift, pressure maintenance and secondary recovery methods; and (d) offering characterizations of both the optimal path of production (15), and the optimal path of investment (16), which include explicit statements of user costs.

The model presented in this work allows the analytical development of arguments suggested in the works of both Davidson and McDonald (1971).[7] In addition, the work of Davidson is extended by allowing variation in production and investment rates over time.

---

[7] Davidson defines three kinds of user costs associated with oil recovery. The first is the user cost associated with all raw materials; i.e., "the more used today, *ceteris paribus*, the less will be available tomorrow" (p. 91). Using the notation of the present paper, this user cost is simply $\lambda\beta_T$ of (15). Davidson's second user cost is associated with rate sensitivity of oil reservoirs, which is captured in (15) by $-\lambda\beta_T(\partial X/\partial u_{jt})$ and by the next to the last term in the right-hand side of (15). Finally, Davidson observes that if a reservoir is controlled by more than one firm, a "cost of the rule of capture" will be incurred (p. 94), which in the present model is essentially measured by the last term in the right-hand side of (15). McDonald (1971, p. 84) states a marginal condition for the optimum rate of unrestrained production by a firm, and identifies terms in his equation as marginal user costs (pp. 98 and 99). These user costs correspond to those defined by Davidson.

Smith discusses crude oil production within the context of a general model which attempts to provide "... a unified theory of production from natural resources" (p. 409). The present work extends that of Smith by providing simultaneously determined characterizations of the optimal paths of production and investment, and by allowing both the path dependence of ultimate recovery, and the use of artificial lift and pressure maintenance.[8]

Kuller presents an optimal control model of oil production for a firm. This analysis is extended in the present model by the explicit treatment of investment.

V. Final Observations

In the above sections an attempt has been made to provide a model for petroleum reservoir management based, as closely as possible, upon a characterization of variables which allows empirical applications using (for physical relations) the current estimation techniques of petroleum engineers. The following two observations are made in conclusion.

First, various writers have been concerned with the problem of determining optimal expenditures for exploration in the petroleum industry (see for example, Orris Herfindahl, pp. 133–38; and McDonald (1971), pp. 86–91). While such a topic is beyond the scope of this paper, some possible extensions of this work may be suggested which might provide interesting insights concerning exploration.

Applying the model developed above in Section II to N reservoirs, an estimate of the nation's (optimally determined) petroleum stock χ_t could be derived for any time t, where

$$\chi_t = \sum_{n=1}^{N} X_n(U_n, V_n) \tag{20}$$

The X_n's are the endogenously determined stocks for each reservoir $n=1, \dots, N$ as defined above in (9). Let E_t be expenditures for exploration in t, and $\Gamma(E_t)$ be the expected value of oil reserves discovered as a result of exploration expenditures E_t. If, for the simplest of examples, national benefits from petroleum production are viewed as the simple sum of net incomes from each reservoir—i.e., the sum of equation (10) over N reservoirs—the appropriate criterion function will include the term $-E_t\beta_t$; and the national stock (ignoring exports and imports) will be expressed by

$$\chi_t = \sum_{n=1}^{N} X_n(U_n, V_n) + \Gamma(E_t) \tag{21}$$

Differentiating such a model with respect to E_t, if E_t is nonzero, a term of the following form appears:

$$\beta_t = \Lambda\beta_t \frac{\partial\Gamma}{\partial E_t} \tag{22}$$

In (22), Λ is the marginal scarcity value of the nation's known petroleum reserves (again, optimally determined, see the above development of λ for the individual reservoir); $\partial\Gamma/\partial E_t$ is the expected increase in known reserves from an increment in E. The term β_t, of course, is the present value of a one dollar marginal expenditure for exploration. Optimum expenditures for exploration are thus the level at which marginal costs of exploration expenditures equal the present value of the expected increase in reserves resulting from the marginal expenditure.

This result is intuitively obvious on a priori grounds; not immediately obvious is a method by which the present value of expected stock-increases might be estimated. Two recent works relate to this problem, however. Maurice Allais has suggested a method for providing a "... reasonable forecast of the economic prospects of mining exploration ..." (p. 285).[9] More recently, Russell Uhler and Paul Bradley suggest a model which may be used for "... determining the economic prospects of petroleum exploration ..." (p. 623).

---

[8] The Smith model considers production solely by natural drive. However, a method for including artificial lift and pressure maintenance in the analysis is suggested (p. 414, n. 6). The possible rate dependence of ultimate recovery is recognized, although not incorporated in the model.

Second, in the discussions above, reference has been made to the problem of choosing T, the planning horizon, and some comments are required. In terms of analytical "neatness," this problem is discussed in some detail in other works (see for example, Burt and Cummings, Appendix). In most empirical applications, a convergence of optimal policies to some fixed path u^* is generally expected (see R. Bellman and Burt). If, for whatever reason, recursive techniques are not used, it will generally be necessary to run a sensitivity analysis on an arbitrarily chosen T, and to continue with increasingly larger values of T until the impact of changes in T, vis-à-vis the resulting solution, is sufficiently small to be regarded as "negligible."

APPENDIX

The optimization problem expressed above by equations (10) through (14) may be treated within the framework of Kuhn-Tucker theory (see G. Hadley, pp. 188, 192, 193). Assuming that the functions in (10) through (13) are continuous, and possess first partial derivatives with respect to each argument, the following Lagrangian expression may be formed, where the Lagrangian multipliers are explicitly discounted for the purpose of economic interpretation.

$$
\begin{aligned}
\text{(A1)}\quad L = \sum_{t=1}^{T}\sum_{j=1}^{n}\{p_t u_{jt} - C_{jt}(U_t, V_t, K_{jt})\}\beta_t \\
- \sum_{j=1}^{n}\sum_{k=1}^{q}\Delta_{jk,t+1}\beta_{t-1}\{K_{jk,t+1} \\
- K_{jkt} + D_{jkt}(u_{jt}, v_{jkt}, K_{jkt})\} \\
- \sum_{j=1}^{n}\Psi_{jt}\beta_t\{u_{jt} - F_{jt}(U_t, V_t, K_{jt})\} \\
- \lambda\beta_T\left\{\sum_{\tau=1}^{T}\sum_{j=1}^{n}u_{j\tau} - X(U_T, V_T)\right\} \\
+ \sum_{j=1}^{n}\xi_{jt}\beta_t u_{jt} + \sum_{j=1}^{n}\sum_{k=1}^{q}\sigma_{jkt}\beta_t v_{jkt}
\end{aligned}
$$

Necessary and sufficient conditions for a maximum in (A1) include the following (given in Hadley, pp. 192–93):

$$
\begin{aligned}
\text{(A2)}\quad \frac{\partial L}{\partial u_{jt}} = \left(p_t\beta_t - \sum_{\tau=t}^{T}\sum_{i=1}^{n}\frac{\partial C_{i\tau}}{\partial u_{jt}}\beta_\tau\right) \\
- \sum_{k=1}^{q}\Delta_{jk,t+1}\beta_{t+1}\frac{\partial D_{jkt}}{\partial u_{jt}} \\
- \Psi_{jt}\beta_t + \sum_{\tau=t}^{T}\sum_{i=1}^{n}\Psi_{i\tau}\beta_\tau\frac{\partial F_{i\tau}}{\partial u_{jt}} \\
- \lambda\beta_T\left(1 - \frac{\partial X}{\partial u_{jt}}\right) + \xi_{jt}\beta_t \leqq 0
\end{aligned}
$$

$$
\text{(A2')}\quad \left(\frac{\partial L}{\partial u_{jt}}\right)u_{jt} = 0
$$

$$
j = 1, \ldots, n;\ t = 1, \ldots, T
$$

$$
\begin{aligned}
\text{(A3)}\quad \frac{\partial L}{\partial v_{jkt}} = \left(-\frac{\partial C_{jt}}{\partial v_{jkt}}\beta_t - \sum_{\substack{i=1\\ i\neq j}}^{n}\frac{\partial C_{it}}{\partial v_{jkt}}\beta_t\right. \\
\left. - \sum_{\tau=t+1}^{T}\sum_{i=1}^{n}\frac{\partial C_{i\tau}}{\partial v_{jkt}}\beta_\tau\right) \\
- \Delta_{jk,t+1}\beta_{t+1}\frac{\partial D_{jkt}}{\partial v_{jkt}} \\
+ \sum_{\tau=t}^{T}\sum_{i=1}^{n}\Psi_{i\tau}\beta_\tau\frac{\partial F_{i\tau}}{\partial v_{jkt}} \\
+ \lambda\beta_T\frac{\partial X}{\partial v_{jkt}} + \sigma_{jkt}\beta_t \leqq 0
\end{aligned}
$$

$$
\text{(A3')}\quad \left(\frac{\partial L}{\partial v_{jkt}}\right)v_{jkt} = 0
$$

---

[9] Although fuels are not included in Allais' study, he asserts that his method is applicable to the search for coal and oil.

$j = 1, \ldots, n; k = 1, \ldots, q; \mathrm{t} = 1, \ldots, \mathrm{T}$

$$\text{(A4)} \quad \frac{\partial L}{\partial K_{jk\mathrm{t}}} = -\frac{\partial C_{j\mathrm{t}}}{\partial K_{jk\mathrm{t}}}\beta_{\mathrm{t}} - \Delta_{jk\mathrm{t}}\beta_{\mathrm{t}} + \Delta_{jk,\mathrm{t}+1}\beta_{\mathrm{t}+1}\left(1 - \frac{\partial D_{jk\mathrm{t}}}{\partial K_{jk\mathrm{t}}}\right) + \Psi_{j\mathrm{t}}\beta_{\mathrm{t}}\frac{\partial F_{j\mathrm{t}}}{\partial K_{jk\mathrm{t}}} \leqq 0$$

$$\text{(A4}'\text{)} \quad \left(\frac{\partial L}{\partial K_{jk\mathrm{t}}}\right) K_{jk\mathrm{t}} = 0$$

$j = 1, \ldots, n; k = 1, \ldots, q; \mathrm{t} = 1, \ldots, \mathrm{T}$

$$\text{(A5)} \quad \frac{\partial L}{\partial \Delta_{jk,\mathrm{t}+1}} = -\left\{K_{jk,\mathrm{t}+1} - K_{jk\mathrm{t}} + D_{jk\mathrm{t}}(u_{j\mathrm{t}}, v_{jk\mathrm{t}}, K_{jk\mathrm{t}})\right\} \geqq 0$$

$$\text{(A5}'\text{)} \quad \left(\frac{\partial L}{\partial \Delta_{jk,\mathrm{t}+1}}\right) \Delta_{jk,\mathrm{t}+1} = 0, \qquad \Delta_{jk,\mathrm{t}+1} \gtreqless 0$$

$j = 1, \ldots, n; k = 1, \ldots, q; t = 1, \ldots, \mathrm{T}$

$$\text{(A6)} \quad \frac{\partial L}{\partial \Psi_{j\mathrm{t}}} = -\left\{u_{j\mathrm{t}} - F_{j\mathrm{t}}(U_{\mathrm{t}}, V_{\mathrm{t}}, K_{j\mathrm{t}})\right\} \geqq 0$$

$$\text{(A6}'\text{)} \quad \left(\frac{\partial L}{\partial \Psi_{j\mathrm{t}}}\right) \Psi_{j\mathrm{t}} = 0, \qquad \Psi_{j\mathrm{t}} \geqq 0$$

$j = 1, \ldots, n; \mathrm{t} = 1, \ldots, \mathrm{T}$

$$\text{(A7)} \quad \frac{\partial L}{\partial \lambda} = -\left\{\sum_{\tau=1}^{\mathrm{T}}\sum_{j=1}^{n} u_{j\tau} - X(U_{\mathrm{T}}, V_{\mathrm{T}})\right\} \geqq 0$$

$$\text{(A7}'\text{)} \quad \left(\frac{\partial L}{\partial \lambda}\right)\lambda = 0, \qquad \lambda \geqq 0$$

$$\text{(A8)} \quad u_{j\mathrm{t}} \geqq 0, \qquad u_{j\mathrm{t}}\xi_{j\mathrm{t}} = 0, \qquad \xi_{j\mathrm{t}} \geqq 0,$$

$j = 1, \ldots, n; \mathrm{t} = 1, \ldots, \mathrm{T}$

$$\text{(A9)} \quad v_{jk\mathrm{t}} \geqq 0, \qquad v_{jk\mathrm{t}}\sigma_{jk\mathrm{t}} = 0, \qquad \sigma_{jk\mathrm{t}} \geqq 0,$$

$j = 1, \ldots, n; k = 1, \ldots, q; \mathrm{t} = 1, \ldots, \mathrm{T}$

Let capital stocks $K_{jk\mathrm{t}}$ be positive for all t, $1 \leqq \mathrm{t} \leqq \mathrm{t}^*$, $\mathrm{t}^* \leqq \mathrm{T}$. From (A4′) it follows that:

$$\text{(A10)} \quad \Delta_{jk\mathrm{t}}\beta_{\mathrm{t}} = -\frac{\partial C_{j\mathrm{t}}}{\partial K_{jk\mathrm{t}}}\beta_{\mathrm{t}} + \Psi_{j\mathrm{t}}\beta_{\mathrm{t}}\frac{\partial F_{j\mathrm{t}}}{\partial K_{jk\mathrm{t}}} + \left(1 - \frac{\partial D_{jk\mathrm{t}}}{\partial K_{jk\mathrm{t}}}\right)\Delta_{jk,\mathrm{t}+1}\beta_{\mathrm{t}+1},$$

for all t, $1 \leqq \mathrm{t} \leqq \mathrm{t}^*$

Solving (A10) for $\Delta_{jk,\mathrm{t}+1}$, then $\Delta_{jk,\mathrm{t}+2}, \ldots, \Delta_{jk\mathrm{t}^*}$, and substituting these terms into (A10) results in the following expression:[10]

$$\text{(A11)} \quad \Delta_{jk\mathrm{t}}\beta_{\mathrm{t}} = \sum_{\tau=\mathrm{t}}^{\mathrm{t}^*}\left(-\frac{\partial C_{j\tau}}{\partial K_{jk\tau}} + \Psi_{j\tau}\frac{\partial F_{j\tau}}{\partial K_{jk\tau}}\right)\beta_{\tau} \cdot \prod_{s=\mathrm{t}}^{\tau-1}\left(1 - \frac{\partial D_{jks}}{\partial K_{jks}}\right) + \prod_{s=\mathrm{t}}^{\mathrm{t}^*}\left(1 - \frac{\partial D_{jks}}{\partial K_{jks}}\right)\Delta_{jk,\mathrm{t}^*+1}\beta_{\mathrm{t}^*+1}$$

It is useful for discussions in the text to note that $\Delta_{jk\mathrm{t}}$ is positive, given[11] $\partial C_j/\partial K_{jk} \leqq 0$, by (3); $\Psi_j \geqq 0$, by (A6′), and $\partial F_j/\partial K_{jk} \geqq 0$, by (2).

For the case where firm j ignores all externalities, discussed in the text, the relevant optimization problem is that of maximizing:

$$\text{(A12)} \quad \text{Max} \sum_{\mathrm{t}=1}^{\mathrm{T}}\left\{p_{\mathrm{t}}u_{j\mathrm{t}} - C_{j\mathrm{t}}(u_{j\mathrm{t}}, v_{j\mathrm{t}}, K_{j\mathrm{t}})\right\}\beta_{\mathrm{t}}$$

subject to:

$$\text{(A13)} \quad \begin{aligned} K_{jk,\mathrm{t}+1} &= K_{jk\mathrm{t}} - D_{jk\mathrm{t}}(u_{j\mathrm{t}}, v_{jk\mathrm{t}}, K_{jk\mathrm{t}}), \\ u_{j\mathrm{t}} &\leqq F_{j\mathrm{t}}(K_{j\mathrm{t}}), \\ u_{j\mathrm{t}} &\geqq 0 \end{aligned}$$

---

[10] In (A11) let b_s denote $(1-\partial D_{jks}/\partial K_{jks})$, then

$$\prod_{s=\mathrm{t}}^{\mathrm{t}-1} b_s = 1$$

$$\prod_{s=\mathrm{t}}^{\mathrm{t}} b_s = b_{\mathrm{t}},$$

$$\prod_{s=\mathrm{t}}^{\mathrm{t}+1} b_s = b_{\mathrm{t}}b_{\mathrm{t}+1}, \text{ etc.}$$

[11] A possible exception would arise with $\partial D/\partial K > 1$, a circumstance which would be expected only in a most pathological case.

REFERENCES

R. G. Agarwal, R. Al-Hussainy, and H. J. Ramey, "The Importance of Water Influx in Gas Reservoirs," *Transactions*, AIME, Nov. 1965, *234*, 1336–42.

S. R. Agria, "Special Tax Treatment of Mineral Industries," in A. C. Harberger and M. J. Bailey, eds., *The Taxation of Income from Capital,* Washington 1969.

M. Allais, "Method of Appraising Economic Prospects of Mining Exploration Over Large Territories," *Management Sci.,* July 1957, *3,* 285–345.

A. Baca, *Prediction of the Performance of a Solution Gas Drive Reservoir by Muskat's Equation,* Computer Contribution 8, State Geological Survey, Univ. Kansas, Lawrence 1967.

R. E. Bellman, *Dynamic Programming,* Princeton 1957.

O. R. Burt, "The Economics of Conjunctive Use of Ground and Surface Water," *Hilgardia,* 1964, *32,* 31–111.

——— and R. G. Cummings, "Production and Investment in Natural Resource Industries," *Amer. Econ. Rev.,* Sept. 1970, *60,* 576–90.

N. J. Clark, "A Review of Reservoir Engineering," *World Oil,* May 1951, Part 1, and June 1951, Part 2.

B. C. Craft and M. F. Hawkins, Jr., *Applied Petroleum Reservoir Engineering,* Englewood Cliffs 1959.

R. G. Cummings, "Some Extensions of the Economic Theory of Exhaustible Resources," *Western Econ. J.,* Sept. 1969, *7,* 201–10.

P. Davidson, "Public Policy Problems of the Domestic Crude Oil Industry," *Amer. Econ. Rev.,* Mar. 1963, *53,* 85–108.

R. L. Gordon, "A Reinterpretation of the Pure Theory of Exhaustion," *J. Polit. Econ.,* June 1967, *75,* 274–86.

J. M. Griffin, "The Process Analysis Alternative to Statistical Cost Functions: An Application to Petroleum Refining," *Amer. Econ. Rev.,* Mar. 1972, *62,* 46–56.

G. Hadley, *Nonlinear and Dynamic Programming,* Reading 1964.

A. C. Harberger, "The Taxation of Mineral Industries," in *Federal Tax Policy for Economic Growth and Stability,* Joint Committee on the Economic Report, 84 Cong., 1st sess., Washington 1955, 439–49.

O. C. Herfindahl, "Some Fundamentals of Mineral Economics," *Land Econ.,* May 1955, *31,* 131–38.

A. E. Kahn, "The Depletion Allowance in the Context of Cartelization," *Amer. Econ. Rev.,* June 1964, *54,* 286–314.

R. G. Kuller, "Optimal Exploitation of Resources: The Case of Petroleum," res. pap. no. 43, Univ. Kansas, Lawrence 1972.

S. L. McDonald, *Petroleum Conservation in the United States: An Economic Analysis,* Baltimore 1971.

———, "Percentage Depletion, Expensing of Intangibles, and Petroleum Conservation," in M. Gaffney, ed., *Extractive Resources and Taxation,* Madison 1967.

J. W. McKie and S. L. McDonald, "Petroleum Conservation in Theory and Practice," *Quart. J. Econ.,* Feb. 1962, *76,* 98–121.

T. E. W. Nind, *Principles of Oil Well Production,* New York 1964.

S. J. Pirson, *Oil Reservoir Engineering,* New York 1958.

J. Quirk and V. Smith, "Dynamic Economic Models of Fishing," in A. Scott, ed., *Proceedings of the H. R. MacMillan Fisheries Economics Symposium,* Vancouver 1969.

A. Scott, "Notes on User Costs," *Econ. J.,* June 1953, *63,* 368–84.

———, "The Theory of the Mine Under Conditions of Certainty," in M. Gaffney, ed., *Extractive Resources and Taxation,* Madison 1967.

V. Smith, "The Economics of Production from Natural Resources," *Amer. Econ. Rev.,* June 1968, *58,* 409–31.

P. O. Steiner, "Percentage Depletion and Resource Allocation," in *Tax Revision Compendium,* House Committee on Ways and Means, 86 Cong., 1st sess., Washington, 1959, *2,* 949–66.

R. S. Uhler and P. G. Bradley, "A Stochastic Model for Determining the Economic Prospects of Petroleum Exploration Over Large Regions," *J. Amer. Statist. Assn.,* June 1970, *65,* 623–30.

L. C. Uren, *Petroleum Production Engineering —Oil Field Exploitation,* New York 1953.

Can Equalization of Opportunity Reduce Social Mobility?

By John Conlisk*

> "Greater wealth, health, freedom, fairness, and educational opportunity are *not* going to give us the egalitarian society of our philosophical heritage. It will instead give us a society sharply graduated, with ever greater innate separation between the top and the bottom, and ever more uniformity within families as far as inherited abilities are concerned." . . . "What is most troubling about this prospect is that the growth of a virtually hereditary meritocracy will arise out of the successful realization of contemporary political and social goals." . . . " . . . improving the environment raises the heritability. The higher the heritability, the closer will human society approach a virtual caste system, with families sustaining their position on the social ladder from generation to generation as parents and children are more nearly alike . . ."
>
> *Richard Herrnstein*

An important area of overlap between psychology and economics concerns the nature-nurture issue in the intergenerational transmission of intelligence, income, and other characteristics. Psychologists are likely to stress parent-child intelligence effects and discuss income (or socioeconomic status or some such concept) as a related issue; economists are likely to reverse the stress. This paper presents an intergenerational income distribution model, stressing the nature-nurture issue, the relation between IQ and economic success, and the long-run distributional impact of equal opportunity policy.

One of the stimuli for writing this paper may be of interest. The quotes above are from an *Atlantic Monthly* article by Harvard psychologist Richard Herrnstein. The article was presumably widely read, since it was the cover article and received advance publicity through *Newsweek* and *Time* magazines. After a layman's introduction to the usefulness of IQ testing, Herrnstein presents some controversial arguments from which the quotes are taken. To the title of this paper, Herrnstein answers that equalizing opportunity not only can, but will reduce social mobility. At first, it seemed to me that this assertion could be easily "disproved" by signing a certain derivative in the model below. However, discussions with a psychologist suggested that Hernstein was in effect concerned with a different derivative, whose sign provided an easy "proof' of the assertions.

I. The Model

The mathematical formulation is a simple system of three difference equations subjected to additive shock terms. The variables are defined on successive generations of a single family. The distribution of variables over all families in society is assumed to be generated by independent drawings on the shock terms. The equations are

$$(1)\qquad Y_t = \alpha_0 + \alpha_1 I_t + \alpha_2 Y_{t-1} + u_{Yt}$$

$$(2)\qquad I_t = \beta_0 + \beta_1 G_t + \beta_2 Y_{t-1} + u_{It}$$

$$(3)\qquad G_t = \gamma_0 + \gamma G_{t-1} + u_{Gt}$$

Here Y_t, I_t, and G_t are the income, IQ, and genetic IQ potential of the t-th

---

[*] Associate professor of economics, University of California, San Diego. The research was supported by National Science Foundation grant GS-3201. For helpful suggestions, special thanks are due to Marjorie Honzik, Richard Schmalensee, the managing editor, and an anonymous referee.

generation of the typical family. The Greek letters are parameters. The u's are random shock terms assumed to have zero means, to have the same variance matrix at each generation t, and to be serially uncorrelated. These assumptions are adequate to determine the variances and covariances among (Y_t, I_t, G_t) and $(Y_{t-1}, I_{t-1}, G_{t-1})$ for a given family tree at each generation t. Given the assumption that all family trees obey the same model (with independent shock term drawings), these variances and covariances will, for a large population, be virtually identical to sample variances and covariances computed over all family trees in society. Thus, by analyzing the equilibrium values of the variances and covariances, we can analyze the equilibrium income dispersion in society, the equilibrium income correlation between generations, and similar questions. The approach has been discussed in my 1969 paper.

If humans were unisexual and each person had exactly one offspring, then each family tree would be a straight line (no branching) and the interpretations of variables would be neat. As it is, some convention must be presumed which associates with a given family exactly one previous generation, or parental, family. That is, once (Y_t, I_t, G_t) is measured on a typical family, some convention is needed to decide which parental family to measure $(Y_{t-1}, I_{t-1}, G_{t-1})$ on. A simple convention would be to follow the male, or the female, line. A possibly more sensible convention would start with a set of rules for determining a family's economic "head." Then the straight line family tree needed by the model could be traced through family heads—family head, head parent of family head, head parent of head parent of family head, and so on. Also needed are conventions about which family members to define (Y_t, I_t, G_t) on. Though not the only choice, the following discussion assumes the IQ variables I_t and G_t are measured on the family head, while Y_t includes all family members' income.

Equation (1) states that a family's income is determined, subject to random shock, by the family head's IQ and by the family income of the head's parents. Here "income" may be used loosely. It might be interpreted as any of a number of general success measures—dollar income, occupational status, schooling attained, and so on. Since the time period is a generation, Y_t should be defined as lifetime income. Furthermore, since equilibrium variances and covariances are the objects of interest here, Y_t is assumed to be measured as a trendless index of some sort (absolute income as a ratio to society's average perhaps). Equations (2) and (3) state that the head's IQ is determined by genetic and environmental factors. The genetic component G_t is determined by equation (3) as the sum of a systematic inheritance from the head's parent and a random shock. Equation (2) states that the head's genetic IQ potential (G_t) is translated into actual IQ (I_t) subject to a systematic environmental effect determined by the family income Y_{t-1} of the head's parents, and by a random environmental effect u_{It} which captures omitted variables.

The apparent n-equation generalization of the model, discussed in the Appendix, is straightforward. The simplifications going into the three equation form (1)–(3) are deliberate. Even slightly greater complication would prevent explicit parametric statement of equilibrium magnitudes below (see the Appendix). Nonetheless, the model as it stands does allow a test of the logic of certain important propositions, and does allow plausible numerical illustration.

The following parameter restrictions, some of which are units normalizations, will be assumed. Here σ_Y^2, σ_I^2, σ_G^2, σ_{YI}, σ_{YG}, and σ_{IG} denote the variances and covariances of u.

$$(4)\quad \alpha_0=\beta_0=\gamma_0=0 \quad \sigma_G^2=1 \quad \alpha_1=\beta_1=1$$

$$0\leq \alpha_2,\ \beta_2,\ \alpha_2+\beta_2<1 \quad 0<\gamma<1$$

$$\sigma_{YG}=\sigma_{IG}=0$$

Since Y_t, I_t, and G_t are all index numbers, it may be assumed without loss of generality that their origin is chosen to make their equilibrium values zero; the restriction $\alpha_0=\beta_0=\gamma_0=0$ accomplishes this. Furthermore, it may be assumed without loss of generality that the genetic IQ index G_t is multiplicatively scaled to give the shock term u_{Gt} unit variance; hence the restriction $\sigma_G^2=1$. It is assumed that IQ contributes positively to income after controlling on parental income (so $\alpha_1>0$); that there is a genetic component to IQ determination (so $\beta_1>0$); and that the parent-child genetic intelligence correlation is positive (so $\gamma>0$). Empirical bases for the assumptions α_1, β_1, $\gamma>0$ are sketched in the following two paragraphs. Given that α_1, $\beta_1>0$, there is no loss of generality in assuming that Y_t and I_t are multiplicatively scaled so that $\alpha_1=\beta_1=1$. So the substantive assumption in $\alpha_1=\beta_1=1$ is that α_1, $\beta_1>0$. The assumptions $\alpha_2\geq 0$ and $\beta_2\geq 0$ state that it can't hurt in terms of income and intelligence to have more affluent parents, all else equal. The assumptions $\alpha_2+\beta_2<1$ and $\gamma<1$ are stability conditions (see the Appendix). The determination of the genetic variable G_t for a person takes place at his conception; and the random term u_{Gt} is an accident of conception. It seems safe to assume that u_{Gt} is uncorrelated with u_{Yt} and u_{It}, which mainly involve different forces occurring much later in time. Hence $\sigma_{YG}=\sigma_{IG}=0$ is assumed.

The assertion $\beta_1>0$ (normalized here to $\beta_1=1$) and $\gamma>0$ amount to saying that genetic factors do contribute to the variance of IQ. Psychologists appear to have strong evidence for these assertions in the form of sample correlations between pairs of individuals with differing degrees of blood relationship and differing similarities of environmental background. For example, the IQ correlations of (genetically) identical twins separated from infancy run much higher (about .75) than for unrelated children raised together (about .25), which suggests a genetic contribution to the variance of IQ. Such correlation evidence can be used in genetic models to estimate the "heritability" of IQ—a zero-to-one measure of the relative importance of the genetic contribution to IQ's variance. In terms of equation (2), heritability roughly translates as the ratio of the variance of the G_t term to the total variance of I_t. The IQ heritability estimates run about .80, suggesting that the genetic contribution is very important. Harold Jones and, more recently, Arthur Jensen present lengthy review articles on this type of evidence. (The Jensen article is the one which stimulated a storm of controversy concerning compensatory education and racial IQ differentials, though that controversy is not to the point here.)

The assertion $\alpha_1>0$ (normalized here to $\alpha_1=1$) says that IQ has a positive partial effect on family income Y_t, controlling on parent's income Y_{t-1}. The most direct type of evidence—that based on direct observations of the triplet (Y_t, I_t, Y_{t-1})—seems very hard to find. My 1971 paper presents such evidence, though based on a small sample; IQ does have the asserted effect, operating through education. Measuring indirectly, a partial correlation of nearly .50 between Y_t and I_t can be pieced together by the formula relating this partial correlation to simple correlations from many separate studies in the literature: about .55 simple correlations between occupational status Y_t and IQ (I_t), (studies briefly reviewed in Jensen, p. 15); about .40 between son's IQ (I_t) and father's status Y_{t-1} (studies reviewed in Jensen,

$$\text{var}\,(Y_t) = [\sigma^2 + \lambda(2\mu - 1)]/[1 - (\alpha_2 + \beta_2)^2] \tag{5}$$

$$\text{var}\,(I_t) = \sigma_I^2 + \lambda + \beta_2^2\,\text{var}\,(Y_t) + 2\beta_2\lambda\mu\gamma \tag{6}$$

$$\text{var}\,(G_t) = \lambda \tag{7}$$

$$r(Y_t, Y_{t-1}) = \alpha_2 + \beta_2 + \lambda\mu\gamma/\text{var}\,(Y_t) \tag{8}$$

$$r(I_t, I_{t-1}) = \frac{\lambda\gamma + \beta_2[\sigma_I^2 + \sigma_{YI} + \mu(2\lambda - 1) + \beta_2\,\text{var}\,(Y_t) r(Y_t, Y_{t-1})]}{\text{var}\,(I_t)} \tag{9}$$

$$r(G_t, G_{t-1}) = \gamma \tag{10}$$

$$r(Y_t, I_t) = \frac{\sigma_I^2 + \sigma_{YI} + \lambda\mu + \beta_2\,\text{var}\,(Y_t) r(Y_t, Y_{t-1})}{[\text{var}\,(Y_t)\,\text{var}\,(I_t)]^{1/2}} \tag{11}$$

$$r(Y_t, G_t) = \mu[\lambda/\text{var}\,(Y_t)]^{1/2} \tag{12}$$

$$r(I_t, G_t) = (1 + \mu\gamma\beta_2)[\lambda/\text{var}\,(I_t)]^{1/2} \tag{13}$$

pp. 74–77); and about .40 between son's status Y_t and father's status Y_{t-1} (see Peter Blau and Otis Duncan, p. 110), which is about the same as father-son years of schooling correlations (see Beverly Duncan). Confidence in the assumption $\alpha_1 > 0$ would also follow from general confidence in (i) the many psychologists' studies of IQ's independent contribution to predictions of educational success, and (ii) the many economists' studies of the contribution of educational to economic success.

The *equilibrium* variances and selected *equilibrium* correlation coefficients implied by the model are shown in equations (5) through (13), (see the Appendix for derivations). The parameter transforms $\lambda = (1-\gamma^2)^{-1}$, $\mu = [1-\gamma(\alpha_2+\beta_2)]^{-1}$, and $\sigma^2 = \sigma_Y^2 + \sigma_I^2 + 2\sigma_{YI}$ have been used to simplify ($\lambda, \mu > 1$ and $\sigma^2 > 0$). The units normalizations $\alpha_1 = \beta_1 = \sigma_G^2 = 1$ and the assumptions $\sigma_{YG} = \sigma_{IG} = 0$ have also been used to simplify.

II. Impacts of Equal Opportunity Policy

Hypotheses about the impact of successful policies to equalize opportunity can be stated in terms of sensitivities of the equilibrium magnitudes (5)–(13) to parameter changes. The size of the parent-child income correlation $r(Y_t, Y_{t-1})$ is a measure of the degree of social immobility in a society described by the model. In these terms, the question posed by the title of this paper is whether equalizing opportunity can increase $r(Y_t, Y_{t-1})$, that is, increase the extent to which a family's income is determined by the income of the head's parents.

If "equalizing opportunity" means reducing the special advantages of being born to affluence and the special disadvantages of being born to poverty, then equalizing opportunity might be interpreted as reducing α_2 or β_2 in the model. The derivative of interest is the following (since α_2 and β_2 appear only as the sum $\alpha_2+\beta_2$ in the expression for $r(Y_t, Y_{t-1})$).

$$\frac{\partial r(Y_t, Y_{t-1})}{\partial(\alpha_2+\beta_2)} = \frac{\sigma^2\{\sigma^2+2\lambda\mu+\mu^2(\lambda-1)[1-(\alpha_2+\beta_2)^2]\}+\lambda\mu^2}{[\sigma^2+\lambda(2\mu-1)]^2} \tag{14}$$

In view of parameter constraints (4), this derivative must be positive; that is, opportunity equalization must increase social mobility. This contradicts the opening quotes, and responds "no" to the title question.

However, there is a second sensible interpretation of equalizing opportunity. Many sources of unequal opportunity are not related to parental affluence or poverty. Reduction of these would be reflected in the model by a reduction in the variances σ_Y^2 and σ_I^2 of the shock terms u_{Yt} and u_{It}. Since σ_Y^2 and σ_I^2 appear in the expression for $r(Y_t, Y_{t-1})$ only within $\sigma^2=\sigma_Y^2+\sigma_I^2+2\sigma_{YI}$ (the variance of $u_{Yt}+u_{It}$), the relevant derivative is

$$(15)\quad \partial r(Y_t, Y_{t-1})/\partial\sigma^2 = -\lambda\mu\gamma[1-(\alpha_2+\beta_2)^2]^{-1}\operatorname{var}(Y_t)^{-2}$$

which must be negative in view of parameter constraints (4); that is, opportunity equalization in the form of a reduction of σ^2 decreases social mobility. This agrees with the opening quotes, and responds "yes" to the title question.

Summarizing, the two types of opportunity equalization—(I) reduction in $\alpha_2+\beta_2$ and (II) reduction in σ^2—will have opposite effects on social mobility. Intuitively, α_2 and β_2 represent systematic parent-child connections in the model, so reducing $\alpha_2+\beta_2$ will reduce the equilibrium parent-child connection measure $r(Y_t, Y_{t-1})$. On the other hand, σ^2 represents random noise in the model, so reducing σ^2 increases the systematic connection measure $r(Y_t, Y_{t-1})$. An example of predominantly type I equalization (reducing $\alpha_2+\beta_2$) might be a compensatory preschool program for children of the poor. An example of predominantly type II equalization might be the replacement of random military conscription by an all-volunteer system. Many equalization policies which come to mind are not clearly of one or the other type, but a mixture. For example, federal subsidization of genetic counseling might reduce the incidence of birth defects at all income levels, thus reducing σ^2; but, to the extent that more affluent persons already received the counseling, there might be a bigger effect at lower income levels, thus reducing $\alpha_2+\beta_2$. For many opportunity equalization policies, therefore, the net effect on social mobility would seem to be an open issue.

The same conclusions apply for an apparent alternate measure of social mobility. In view of the parameter normalizations $\alpha_0=\beta_0=\gamma_0=0$ and $\alpha_1=\beta_1=1$, substitution of (2) and (3) in (1) yields the reduced form equation for Y_t.

$$(16)\quad Y_t = \gamma G_{t-1} + (\alpha_2+\beta_2)Y_{t-1} + (u_{Yt}+u_{It}+u_{Gt})$$

The equilibrium coefficient of determination, call it R_Y^2, of this equation is a reasonable measure of social immobility. In view of the parameter normalization $\sigma_G^2=1$ and the assumptions $\sigma_{YG}=\sigma_{IG}=0$, the error term variance of (16) is $1+\sigma^2$. Hence R_Y^2 is

$$(17)\quad R_Y^2 = 1-(1+\sigma^2)/\operatorname{var}(Y_t)$$

from which the derivatives of R_Y with respect to $\alpha_2+\beta_2$ and σ^2 follow.

$$(18)\quad \partial R_Y/\partial(\alpha_2+\beta_2) = (1/R_Y)(1-R_Y^2)[\lambda\gamma\mu^2+(\alpha_2+\beta_2)\cdot\operatorname{var}(Y_t)]/[\sigma^2+\lambda(2\mu-1)]$$

$$\partial R_Y/\partial\sigma^2 = -[\lambda(2\mu-1)-1]/\{2R_Y[1-(\alpha_2+\beta_2)^2][\operatorname{var}(Y_t)]^2\}$$

In view of parameter constraints (4) (and their implication that $\lambda, \mu>1$), these two derivatives must be positive and negative, respectively. Therefore, the preceding discussion of types I and II equalization apply to the social immobility measure R_Y as well as to the measure $r(Y_t, Y_{t-1})$.

The paradoxical statement that opportunity equalization might reduce social mobility is in a sense a contrived paradox. To many advocates of opportunity equalization, the major goal is a reduction in the income inequality measure var (Y_t), rather than the social immobility measures $r(Y_t, Y_{t-1})$ and R_Y. The relevant derivatives are

$$(19)\quad \partial \operatorname{var}(Y_t)/\partial(\alpha_2+\beta_2) = 2[\lambda\gamma\mu^2 + (\alpha_2+\beta_2)\operatorname{var}(Y_t)] / [1-(\alpha_2+\beta_2)^2]$$

$$\partial \operatorname{var}(Y_t)/\partial\sigma^2 = 1/[1-(\alpha_2+\beta_2)^2]$$

Since both of these magnitudes must be positive, then either type of opportunity equalization will reduce income inequality. The paradox is thus a result of casual verbal equations between income equality and social mobility, whereas the two concepts are distinct (at least as measured here) and can move in opposite directions.

The heritability concept introduced above—the share of genetic factors in the variance of a measure—would best be represented here by var (G_t)/var (I_t) and var (G_t)/var (Y_t), for IQ and income, respectively. Barring eugenic changes in var (G_t), these heritability measures are just inverse measures of inequality in IQ and income. Thus, the statement that opportunity equalization decreases inequality is equivalent to the statement that opportunity equalization increases heritability, though this equivalence is not apparent in some contexts (in the opening quotes for example).

III. Rough Orders of Magnitude

The model, in the simple form given or in much expanded form, would present no unusual estimation problems if data were available. Unfortunately, intergenerational data of the sort needed are not available (G_t is not even measurable). Nevertheless, rough orders of magnitude for the derivatives discussed above, those of $r(Y_t, Y_{t-1})$, R_Y, and var (Y_t) with respect to $\alpha_2+\beta_2$ and σ^2, can be inferred from evidence in the literature. These derivatives can be expressed as functions of the three correlations $r(Y_t, Y_{t-1})$, $r(G_t, G_{t-1})$, and $r(Y_t, G_t)$;[1] and there is evidence in the literature suggesting the following specifications for the three correlations.

$$(20)\quad r(Y_t, Y_{t-1}) = .4 \qquad r(G_t, G_{t-1}) = .5$$

$$.40 \le r(Y_t, G_t) \le .75$$

Table 1 presents values of the derivatives for these specifications. If the magnitudes of the derivatives are judged by the sizes of the elasticities (in parentheses on the table), it appears that type I opportunity equalization (reduction of $\alpha_2+\beta_2$) has a larger impact (in absolute value) on social mobility than type II equalization (reduction in σ^2). This is encouraging since type I equalization increases mobility, whereas type II diminishes it. The elasticities of var (Y_t) suggest a reverse pattern; type II equalization appears to have a larger impact on inequality than type I equalization.

Specifications (20) need justification. Here "income" Y_t may be interpreted as any general economic success measure. Father-son dollar income data for Y_t are not available; but father-son occupational status correlations and years of schooling correlations[2] cluster remarkably closely about the value $r(Y_t, Y_{t-1}) = .4$ specified in (20).

---

[1] Equations (5), (8), (10), (12), (14), (15), (18), and (19) form a system of 10 equations determining the 10 magnitudes $\alpha_2+\beta_2$, γ, σ^2, $\partial r(Y_t, Y_{t+1})/\partial(\alpha_2+\beta_2)$, $\partial r(Y_t, Y_{t-1})/\partial\sigma^2$, $\partial R_Y/\partial(\alpha_2+\beta_2)$, $\partial R_Y/\partial\sigma^2$, ∂var $(Y)_t/\partial(\alpha_2+\beta_2)$, ∂var$(Y_t)/\partial\sigma^2$, and var(Y_t) as functions of the three correlations $r(Y_t, Y_{t-1})$, $r(G_t, G_{t-1})$, and $r(Y_t, G_t)$. Explicit solutions are available, but are excessively tedious to present here.

[2] B. Duncan reports father-son years of schooling correlations of about .4; unpublished calculations yield a similar result for the sample analyzed in Morgan et al. Peter Blau and Otis Duncan, p. 110, report about the same again for father-son occupational status correlations. In each case the sample is large, representative of the United States, and apparently of high quality. Furthermore, the correlations for subsamples by son's age group do not vary much around the .4 benchmark (which lends some credence to the use of equilibrium analysis).

TABLE 1—DERIVATIVES

Assumed Value of $r(Y_t, G_t)$	Derivatives and (in Parentheses) Elasticities of $r(Y_t, Y_{t-1})$ with respect to $\alpha_2+\beta_2$	σ^2	of R_Y with respect to $\alpha_2+\beta_2$	σ^2	of var (Y_t) with respect to $\alpha_2+\beta_2$	σ^2
.75	.92	−.10	.73	−.08	2.34	1.02
	(.32)	(−.28)	(.22)	(−.20)	(.12)	(.43)
.70	.93	−.07	.75	−.06	2.85	1.03
	(.41)	(−.28)	(.29)	(−.21)	(.15)	(.50)
.65	.94	−.05	.77	−.05	3.49	1.05
	(.50)	(−.27)	(.35)	(−.22)	(.19)	(.56)
.60	.95	−.04	.79	−.04	4.29	1.06
	(.57)	(−.25)	(.42)	(−.22)	(.22)	(.62)
.55	.96	−.02	.81	−.03	5.33	1.08
	(.64)	(−.22)	(.49)	(−.22)	(.24)	(.68)
.50	.97	−.02	.84	−.02	6.71	1.09
	(.71)	(−.20)	(.56)	(−.20)	(.27)	(.73)
.45	.97	−.01	.86	−.01	8.56	1.11
	(.77)	(−.17)	(.63)	(−.19)	(.29)	(.78)
.40	.98	−.01	.89	−.01	11.16	1.13
	(.81)	(−.14)	(.69)	(−.16)	(.31)	(.83)

Note: All entries assume $r(Y_t, Y_{t-1}) = .4$ and $r(G_t, G_{t-1}) = .5$

Since the genetic IQ potential G_t is not measurable, there are no reports in the literature on $r(G_t, G_{t-1})$. However, L. Erlenmeyer-Kimling and Lissy Jarvik, in surveying parent-child correlations of measured IQ, suggest a consensus figure of about $r(I_t, I_{t-1}) = .5$. Three reasons may be advanced for speculating that $r(G_t, G_{t-1})$ is about the same, as specified in (20). First, estimates of the heritability var (G_t)/var (I_t) of intelligence run quite high (about .8 as surveyed by Jensen, pp. 46–51), suggesting that G_t and I_t vary closely together. Second, the forces causing a divergence between $r(I_t, I_{t-1})$ and $r(G_t, G_{t-1})$ push in both directions, thus possibly cancelling. In equation (2), the noise term u_{It} tends to make $r(I_t, I_{t-1})$ smaller than $r(G_t, G_{t-1})$, while the systematic term $\beta_2 Y_{t-1}$ tends to make $r(I_t, I_{t-1})$ larger. Third, $r(G_t, G_{t-1}) = .5$ is the theoretical value predicted by very simple genetic models (see Jensen, p. 49).

The specification $.40 \leq r(Y_t, G_t) \leq .75$ in (20) cannot be verified directly from the literature, since G_t is not measurable; but reported values of $r(Y_t, I_t)$ are suggestive proxies. In surveying the literature, Jensen, pp. 15–16, reports $r(Y_t, I_t)$ values from .42 to .57, where Y_t is measured as occupational status; and Jacob Miner, p. 67, reports $r(Y_t, I_t)$ values from .44 to .75 where Y_t is measured as years of schooling.

Table 2 provides further checks on the plausibility of specifications (20) by presenting the implied values of some other magnitudes. The tabled income heritability var (G_t)/var (Y_t) figures are plausible. There is evidence in the literature (reviewed in Jensen, pp. 58–59) that scholastic achievement has on the order of half the heritability of IQ, presumably because many environmental factors intervene between potential and achievement. For the same reason, one might guess that income will have substantially lower heritability than IQ (where, as noted in the paragraph

TABLE 2—OTHER MAGNITUDES

Assumed Value of $r(Y_t, G_t)$	Heritability var (G_t)/var (Y_t)	σ^2	$1/(1+\sigma^2)$	$\alpha_2+\beta_2$	R_Y
.75	.49	1.15	.47	.14	.21
.70	.41	1.58	.39	.18	.21
.65	.34	2.12	.32	.21	.21
.60	.28	2.81	.26	.24	.21
.55	.23	3.71	.21	.27	.20
.50	.18	4.90	.17	.29	.19
.45	.14	6.52	.13	.32	.19
.40	.11	8.80	.10	.33	.18

Note: All entries assume $r(Y_t, Y_{t-1})=.4$ and $r(G_t, G_{t-1})=.5$

following (4), IQ has estimated heritability of about .8). The heritability figures on Table 2 are consistent with this guess. To interpret the σ^2 figures on Table 2, it is helpful to recall that $1+\sigma^2$ is the variance of the shock terms $u_{Yt}+u_{It}+u_{Gt}$ of the reduced form equation (16) for Y_t. In view of the normalization $\sigma_G^2=1$, the ratio $1/(1+\sigma^2)$ is a measure of the genetic contribution to the variance of the Y_t contemporaneous shock term—a short-run heritability of sorts. The tabled values of $1/(1+\sigma^2)$ run about the same as var (G_t)/var (Y_t). In this sense, one might say that short- and long-run heritability are about the same, which does not appear to violate any intuitive expectations. The tabled values of $\alpha_2+\beta_2$ add slightly to the credibility of specification (20) by not going negative. The tabled values of R_Y^2 are consistent with evidence from the economic literature. Regressions of income on long lists of explanatory variables yield R^2 in the neighborhood of .35. Since R_Y^2 is based on an equation with only two explanatory variables (father's potential IQ and father's income), R_Y^2 might be expected to run substantially less than .35, as it does.

IV. A Redistributive Tax Application

Let Y_t represent dollar income after all existing government tax and spending mechanisms have been accounted for. Now suppose the government adds a linear redistributive income tax, specified so taxes are negative below the mean income and positive above the mean income. Such a specification assures that the net revenue of the tax is zero; it is purely a redistributive tax. If r is the marginal tax rate and Y_t^a is income after this additional tax, then the equation $Y_t^a=rE(Y_t)+(1-r)Y_t$ describes the tax; taking expected values yields the zero net revenue condition $E(Y_t^a)=E(Y_t)$. The entire model naturally rewrites as

$$\begin{aligned} Y_t^a &= rE(Y_t) + (1-r)Y_t \\ Y_t &= I_t + \alpha_2 Y_{t-1}^a + u_{Yt} \\ I_t &= G_t + \beta_2 Y_{t-1}^a + u_{It} \\ G_t &= G_{t-1} + u_{Gt} \end{aligned}$$

where the units normalizations $\alpha_0=\beta_0=\gamma_0=0$ and $\alpha_1=\beta_1=1$ have been used. The new expressions for var (Y_t), $r(Y_t, Y_{t-1})$, and R_Y are the same as the old except that $\alpha_2+\beta_2$ must be replaced by $(1-r)(\alpha_2+\beta_2)$ wherever it appears. Since $(1-r)(\alpha_2+\beta_2)$ now plays the role of $\alpha_2+\beta_2$, it follows from (14), (18), and (19) that the imposition of the redistributive tax both reduces inequality and increases social mobility. To get a notion of magnitude, let specifications (20) hold with $r(Y_t, G_t)$ fixed at .5. Then $\alpha_2+\beta_2=.29$, $\gamma=.5$, and $\sigma^2=4.90$. For these parameter magnitudes, the following measures of mobility and inequality follow for $r=0$ and $r=.15$.

	$r(Y_t, Y_{t-1})$	R_Y^2	var (Y_t)
$r=0$	.40	.19	7.32
$r=.15$	.36	.16	7.05

V. Conclusion

Though the model has been kept simple, the linear form can accommodate a large number of additional variables and still be numerically operational. The data problem cannot easily be overcome so long as the model involves parent-child observations on intelligence variables (especially an unmeasurable one like G_t). However, a useful model of this type could be constructed for more available data. For example, a cross-family sample of data on income, occupation, education, IQ, wealth, family size, and other variables is not hard to collect. Matched data for the head's parents would not be available for all variables; but the head could provide enough of them (parental occupation, education, family size, and others) to allow sensible modelling of intergenerational mechanisms.

As it stands, the model has no provision for the special social immobilities which face various subgroups of society (ethnic minorities for example). A straightforward way to handle such immobilities is to specify different parameter values for the various subgroups. The model would then generate a separate income distribution for each subgroup.

Another natural extension of the model, given recent interest in random coefficient regression models, is to add a (zero mean) shock term to each coefficient of the model as well as to each equation of the model. In the case of n equations (and thus n^2 coefficients), this would add $n^2(n^2+1)/2$ additional parameters to the model—the variances and covariances of the n^2 coefficient shocks. Thus, in the general case, the extension would be expensive in number of parameters. However, one special case may be of interest. In most of the expressions of interest above, the crucial parameters α_2 and β_2 appear only as the sum $\alpha_2+\beta_2$. Suppose this sum is subjected to a serially uncorrelated, additive shock of constant variance; while other coefficients remain unshocked. Then, as the Appendix shows all of the derivatives discussed in Section II retain their signs. That is, relative to the issues raised in this paper, the addition of a shock to $\alpha_2+\beta_2$ leaves the model qualitatively unchanged.

Finally, the model's answer to the title question and to the opening quotes may be summarized. Opportunity equalization may increase or reduce social mobility, depending on the type of equalization. Either way, however, opportunity equalization will reduce inequality. Therefore, the opening quotes seem unreasonably pessimistic. If opportunity equalization reduces inequality, it is hard to see how this equalization will "give us . . . ever greater separation between the top and the bottom." And, even if opportunity equalization does reduce social mobility, a "virtual caste system" seems unlikely. In the extreme case of total opportunity equalization (represented in the model by $\alpha_2=\beta_2=\sigma^2=0$), the immobility measures $r(Y_t$ $Y_{t-1})$ and R_Y would reduce to $r(G_t, G_{t-1})$; and psychological evidence suggests a value for $r(G_t, G_{t-1})$ of about .5 (see Section III). A society where only 25 percent of children's income variance was explained by parents' incomes would not seem well described by virtual caste system.

APPENDIX

Derivation of Equilibrium Expressions

The model of the text is a special case of the following linear difference equation system with additive shock term.

$$\text{(A1)} \qquad y_t = Ay_{t-1} + u_t, \qquad E(u_t) = 0,$$

$$E(u_t u_s') = \begin{cases} U & \text{if t} = s \\ 0 & \text{if t} \neq s \end{cases}$$

Here A is an (n, n) constant matrix; U is an (n, n) nonnegative definite constant matrix; and y_t and u_t are $(n, 1)$ random variable vectors at time t. The system (A1) is defined on a typical family tree. The elements of y_t are measurements on n characteristics of the t-th generation of the family (income, IQ, and so on). The distribution of these characteristics over all families is generated by independent drawings on the shock vector u_t.

Corresponding to the stochastic system (A1) are two nonstochastic systems, one n equation system for the mean of y_t, call it μ_t, and one n^2 equation system for the variance matrix of y_t, call it S_t:

$$\text{(A2)} \qquad \mu_t = A\mu_{t-1}$$

$$\text{(A3)} \qquad S_t^* = U^* + (A \otimes A)S_{t-1}^*$$

Here an asterisk indicates the "stacking" operation; for example, S_t^* is the $(n^2, 1)$ vector gotten by stacking the columns of S_t (first column on top, second column next, and so on).

Taking expectations of (A1) leads immediately to (A2). Taking the variance matrix of (A1) leads to $S_t = AS_{t-1}A' + U$, and (A3) is simply the stack of this. The condition that the eigenvalues of A be less than one in modulus is sufficient for the existence of unique equilibria $\mu = 0$ and $S^* = (I - A \otimes A)^{-1}U^*$ for (A2) and (A3), and is necessary and sufficient for their global stability. (The coincidence of asymptotic properties for (A2) and (A3) follows from the theorem that the n^2 eigenvalues of the Kronecker product $A \otimes A$ equal the n^2 pairwise products of the eigenvalues of A.) Letting R and R_{-1} be the equilibrium correlation matrices of y_t with itself and with y_{t-1}, respectively, the relevant formulas for the magnitudes in the text are

$$\text{(A4)} \qquad S^* = (I - A \otimes A)^{-1}U^*,$$
$$R = WSW, \qquad R_{-1} = WASW$$

Here $W = \text{diag}\,(s_{11}^{-1/2}, \ldots, s_{nn}^{-1/2})$, where the s_{ii} are the diagonal elements of S.

The three-equation model (1)–(3) in the text can be reduced to a two-equation model in Y_t and G_t by using (2) to substitute out I_t. The resulting system is of the form (A1) with $y_t = (Y_t, G_t)'$ and

$$\text{(A5)} \qquad A = \begin{pmatrix} \alpha_2 + \beta_2 & \gamma \\ 0 & \gamma \end{pmatrix}$$

$$U = \begin{pmatrix} 1 + \sigma^2 & 1 \\ 1 & 1 \end{pmatrix}$$

Here the units normalizations $\alpha_1 = \beta_1 = \sigma_G^2 = 1$ and the assumptions $\sigma_{YG} = \sigma_{IG} = 0$ have been used to simplify. The eigenvalues $\alpha_2 + \beta_2$ and γ of A lie between zero and one by assumptions (4); hence the model has a unique globally stable equilibrium. Plugging (A5) into (A4) and manipulating yields the equilibrium expressions var (Y_t) and var (G_t) as the diagonal elements of S, $r(Y_t, G_t)$ as the off-diagonal of R, and $r(Y_t, Y_{t-1})$ and $r(G_t, G_{t-1})$ as the diagonals of R_{-1}. The equilibrium expressions involving I_t can then be derived using equation (2).

Even though the condensed system of (A5) is only of size $n = 2$, the matrix to be inverted in (A4) is of size $n^2 = 4$, which would already be too large for explicit inversion in terms of parameters if A were not triangular. This explains the comment in the text that even slightly greater complexity in the model would prevent explicit parametric statement of equilibrium magnitudes.

Now suppose equation (A1) is expanded to allow random coefficients. Specifically, let

$$\text{(A6)} \qquad y_t = (A + Z_t)y_{t-1} + u_t$$

where Z_t is an (n, n) matrix of zero mean coefficient shocks at time t. Let each element of Z_t be serially uncorrelated with itself, with all other elements of Z_t, and with all elements of u_t. And let each element of Z_t have constant variance, and constant covariance with all other elements of Z_t and all elements of u_t. For this expanded model, my 1972 paper showed that the equilibrium variance matrix of y_t, call it S_+, is given in stacked form by

$$\text{(A7)} \qquad S_+^* = (I - A \otimes A - C)^{-1}U^*$$

Here C is (n^2, n^2) constant matrix displaying the variances and covariances of the elements of Z_t in the form $C=E(Z_t \otimes Z_t)$ (which is not usual variance matrix form). The necessary and sufficient global stability condition for the equilibrium (A7) is that every eigenvalue of $A \otimes A + C$ have modulus less than one. The formulas for the contemporaneous and one lag correlation matrices, call them R_+ and $(R_+)_{-1}$, are similar to those in (A4).

$$\text{(A8)} \qquad R_+ = W_+ S_+ W_+ \quad \text{and} \qquad (R_+)_{-1} = W_+ A S_+ W_+$$

where $W_+ = \text{diag}\ (s_{+11}^{-1/2}, \ldots, s_{+nn}^{-1/2})$. In the special case when a_{11} is the only element of A with a positive variance shock term, C takes the form $c_{11}E_{11}$, where E_{11} is an (n^2, n^2) matrix of all zeros except for a one in the upper left position. Then (A7) manipulates to the form

$$\text{(A9)} \quad S_+^* = \{I + [c_{11}/(1 - a_{11}^2 - c_{11})]E_{11}\} \cdot (I - A \otimes A)^{-1} U^* = \{I + [c_{11}/(1 - a_{11}^2 - c_{11})]E_{11}\} S^*$$

where, as above, $S^* = (I - A \otimes A)^{-1} U^*$ (the value of S^* when $C=0$).

The formulas of the last paragraph may be applied to the special case discussed in the concluding section of the text. In this special case, $n=2$; A and U are given by (A5); and C takes the form $C=c_{11}E_{11}$, where c_{11} is the variance of the shock term added to $a_{11} = \alpha_2 + \beta_2$. To assure stability, the parameter restrictions of the text must be strengthened to include the restriction $(\alpha_2+\beta_2)^2 + c_{11} < 1$. Plugging into formulas (A7), (A8), and (A9) yields the equilibrium magnitudes $\text{var}\ (Y_t)_+$, $r(Y_t, Y_{t-1})_+$, and $(R_Y^2)_+$ for the expanded model.

$$\begin{aligned} \text{(A10)} \quad & \text{var}\ (Y_t)_+ \\ & = \text{var}\ (Y_t)[1 - (\alpha_2 + \beta_2)^2] \\ & \quad /[1 - (\alpha_2 + \beta_2)^2 - c_{11}] \\ & r(Y_t, Y_{t-1})_+ \\ & = \{r(Y_t, Y_{t-1})[1 - (\alpha_2 + \beta_2)^2 - c_{11}] \\ & \quad + c_{11}(\alpha_2 + \beta_2)\}/[1 - (\alpha_2 + \beta_2)^2] \\ (R_Y^2)_+ & = 1 - (1 + \sigma^2)/\text{var}\ (Y_t)_+ \\ & = 1 - (1 - R_Y^2)[1 - (\alpha_2 + \beta_2)^2 - c_{11}] \\ & \quad /[1 - (\alpha_2 + \beta_2)^2] \end{aligned}$$

where $\text{var}\ (Y_t)$, $r(Y_t, Y_{t-1})$, and R_Y^2 are the expressions given by (5), (8), and (17) in the text. The derivatives of each line of (A10) with respect to $\alpha_2+\beta_2$ and σ^2 take the same signs as the corresponding derivatives in the text for the nonexpanded model ((14), (15), (18), (19)).

REFERENCES

P. M. Blau and O. D. Duncan, *The American Occupational Structure*, New York 1967.

J. Conlisk, "An Approach to the Theory of Inequality in the Size Distribution of Income," *Western Econ. J.*, June 1969, *7*, 180–6.

———, "A Bit of Evidence on the Income-Education-Ability Interrelation," *J. Hum. Resources*, summer 1971, *6*, 358–62.

———, "Stochastic Stability of Random Coefficient Models," mimeo. 1972.

B. Duncan, "Education and Social Background," *Amer. J. Sociology*, Jan. 1967, *72*, 363–72.

L. Erlenmeyer-Kimling and L. F. Jarvik, "Genetics and Intelligence: A Review," *Science*, Dec. 13, 1963, *142*, 1477–8.

R. Herrnstein, "I.Q.," *Atlantic*, Sept. 1971, *228*, 43–64.

A. R. Jensen, "How Much Can We Boost IQ and Scholastic Achievement?" *Harvard Educ. Rev.*, winter 1969, *39*, 1–123.

H. E. Jones, "The Environment and Mental Development," in L. Carmichael, ed., *Manual of Child Psychology*, 2d ed., New York 1954.

J. B. Miner, *Intelligence in the United States*, New York 1957.

J. N. Morgan, M. H. David, W. J. Cohen, and H. E. Brazer, *Income and Welfare in the United States*, New York 1962.

Newsweek, Aug. 23, 1971, p. 57.

Time, Aug. 23, 1971, p. 33.

Toward a Theory of Multiple Customs Unions

By SUKESH K. GHOSH*

In a recent article in this *Review*, Sven Arndt extended the popular three-country analysis of customs unions to four countries consisting of two member countries of the union, on the one hand, and two countries of the rest of the world, on the other. Based on this extension, he demonstrated the possibility of opposite welfare effects of the union on the nonmember countries so that one of the conclusions, in his words, was that "... it may be rational policy for some countries [*outside the union*] to push for the establishment of a given customs union, but nevertheless to refrain from joining that union"[1] (p. 117).

However, this result which clearly applies to one customs union will hold even when the model is more general to include more than one customs union. In fact, taking a world of two customs unions and two outside countries, we shall show in this paper that the above result is true. What is more important, the multiple customs unions issue will open up some entirely new possibilities: These are in the form of contrasts in the welfare and terms of trade effects on the *unions themselves* (let alone the members of a given union), possibilities that are characteristically obliterated in the traditional[2] single union analysis.

The following pages will be devoted to the exploration of such and other possibilities in a systematic way. The central method used will be a combination of excess offer curves introduced by Vanek, and aggregate offer curves derived[3] by Gary Becker and George Elliot, recently used by Arndt, and modified in this paper to ensure an efficient[4] integration of individual country offer curves.

I. The Model

First of all, we shall state the basic assumptions: full employment of resources, inelastic factor supply, absence of transportation costs, homogeneous and mobile commodities, (internationally) immobile factors of production, identical tastes and preferences of individuals within each country, respectively concave and convex (to the origin) production transformation curves and community indifference curves, regulatory tariffs. The last assumption along with a supplementary one that the tariff revenue be distributed as income subsidies to the community will eliminate the (otherwise realistic) possibility of changes in income distribution and of intersecting reciprocal demand curves.

The structure of the model can be simply described as follows. We imagine the world to consist of six countries,[5] three of which

---

* Associate professor of economics, University of Waterloo, Canada. I gratefully acknowledge the help received from Sven Arndt, Robert Mundell, and a referee of this *Review*. However, I alone am responsible for errors and imperfections.

[1] Preference for nonmembership is also on grounds of increased markets and growth potential outside the union which accrues to the gaining nonmember country.

[2] See for instance, Jacob Viner, Richard Lipsey, Michael Michaely, Jaroslav Vanek, C. A. Cooper and Benton Massell.

[3] As Arndt notes, p. 109, fn. 3, aggregate offer curves were implied by Frank Graham.

[4] It should be stated here that the modification used in this paper will not affect any of Arndt's conclusions, for they were obtained from sections of the aggregate offer curves basically unaffected by the changes I introduce. However the changes will prove crucial for the conclusion of this paper.

[5] Each of these countries will be taken to be incompletely specializing in X and Y.

to be called A_1, A_2, and A_3 represent a group of Y (commodity) exporting and the remaining three, B_1, B_2, and B_3 a group of X (commodity) exporting countries over a given range of the terms of trade. In order not to complicate and blur the main issues, we will posit that changes in the terms of trade never go beyond that range so that the nature of the comparative advantage of each country will remain unaffected throughout the analysis. Also, to bring (later) into sharp relief possible trade diverting and trade creating effects of the formation of customs unions, we shall consider a relative ranking of the following types of the three countries in each group: $A_1 < A_2 < A_3$; $B_1 < B_2 < B_3$. The ranking proposed is on the basis of both the cost of producing the commodity in which a country in a given group has a comparative advantage and the *ad valorem* tariff that it levies on its imports. However, a simple calculation will make one wonder about many other cost and tariff patterns. Such patterns, it may be noted, will follow the analysis, to be undertaken below, of the central pattern just described.

In the spirit of the above paragraphs, we will broadly consider the following two customs unions,[6] each of which consists of two countries, the residual countries (A_1 and B_1) comprising the nonunionized countries of the rest of the world:

$$\text{Union } u_1\text{: } A_2\, U B_2$$

$$\text{Union } u_2\text{: } A_3\, U B_3$$

None of these unions is assumed to be small so that for neither union the world terms of trade is given.

We also propose a composite union:

$$\text{Union } U\text{: } A_2\, U A_3\, U B_2\, U B_3$$

to be used as a bench mark union for comparative purposes.

II. The Preunion Situation

The preunion situation offer curves, each adjusted for the respective country's *ad valorem* tariff are shown as OA_1', OA_2', and OA_3' for the A countries, and OB_1', OB_2', and OB_3' for the B countries in Figure 1, the former curves exhibiting relative offers of the Y (shown on the Y-axis) for the X commodity (shown on the X-axis) and the latter curves the offers of X for Y.

By a modification of the well-known method of radial sum of individual offer curves, the group offer curves $OILW'_A$ and $OJKW'_B$, respectively, for the A and B groups of countries are derived. This type of group offer curves was also derived by Arndt except for their linear parts as, for instance, IL and JK for the A group and B group offer curves, respectively. The linearity is the modification proposed here which was perhaps overlooked by the earlier authors,[7] most recently by Arndt. It is derived, in detail, in the Appendix.

In the preunion situation, the world terms of trade are given by the slope of the line OT_t, where T_t is the point of intersection of the two (modified) group offer curves. The internal prices in the A countries are however given by the slopes of the lines UQ, TP, and MS, being tangent to the A_1, A_2, and A_3 country's trade indifference curves at points U, T, and S, respectively. At these points, OT_t intersects the respective country's tariff inclusive offer curve. Notice[8] that the tariff rate (OQ/QR) in country A_1 is less than that (OP/PV) in A_2 which is less than that (OM/MN) in A_3. Similarly, the internal price lines in the B countries may be drawn (not shown in the diagram) consistent

---

[6] Later (in fn. 32) we will indicate the terms of trade and welfare effects of differently constituted unions such as Union u_3:A_2UB_3; Union u_4:A_3UB_2; each of which represents a union of high and low cost producing countries.

[7] Gary Becker and George Elliot, for example.

[8] This is the assumed tariff relationship among the A countries.

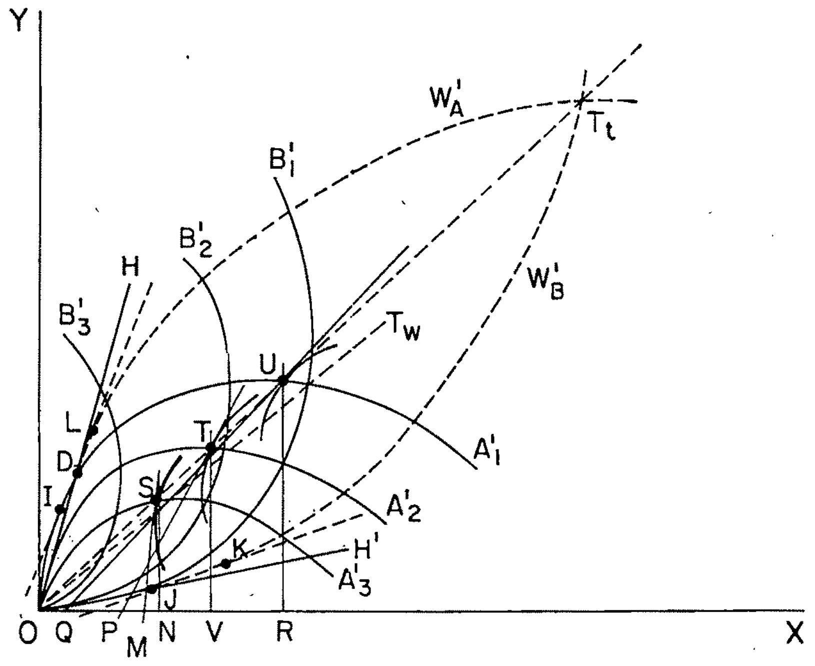


FIGURE 1

with the assumed relationship among their *ad valorem* tariffs.

III. Postunion (u_1 and u_2) Situation

The basic apparatus shown in Figure 1 is transferred to Figure 2 with the preunion situation world terms of trade OT_t directly extended to this figure. The observations below refer to Figure 2 unless otherwise mentioned.

Assume that each union removes all trade barriers between its members, but it pursues a policy of nonprohibitive tariff towards the outside world. To simplify matters, assume initially that the tariff imposed by each union on its imports is the same. Assume[9] also that, for a union, one member country's trade is exclusively with its partner and that the latter only carries out trade with the outside countries in respect of the excess supply of the commodity in which it commands a comparative advantage, and that such excess supplies are in two different commodities for the two unions. To be exact, let us say country A_2 and B_3 of union u_1 and u_2 are the countries with eventual excess supplies in commodity Y and X, respectively.

Under the above assumptions, we construct, following Vanek (ch. 18), the excess offer curves designated as $OE_{2,f}$ and $OE_{3,f}$: The former is derived from OA_2 and OB_2, and the latter from OB_3 and OA_3, one pair for a union. Taking into account the import duties of each union, we then generate the excess offer curves $OE_{2,t}$ and $OE_{3,t}$ which when integrated[10] with the tariff-inclusive offer curves OA_1' and OB_1' of country A_1 and B_1, respectively, of the outer world will yield the aggregate offer

---

[9] This assumption too is used for simplicity's sake, an example of which can be found in the theoretical developments in Vanek, ch. 18. However, simplicity, perhaps, reduces the scope of applicability of the results to be derived.

[10] By the method of modified radial sums described before.

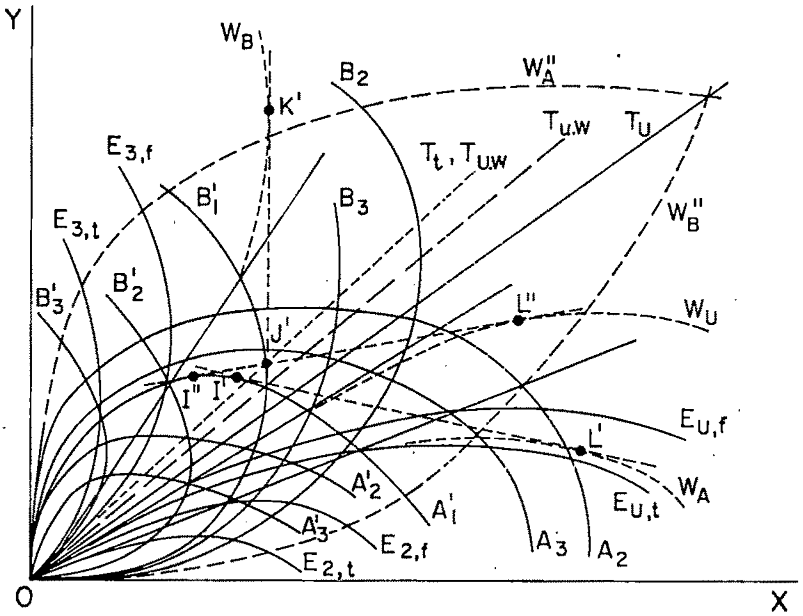


FIGURE 2

curves $OI'L'W_A$ and $OJ'K'W_B$. The world terms of trade $OT_{u.w}$ will clearly settle at the intersection of the last mentioned offer curves.

We shall now dwell upon some of the important features of Figure 1. It is clear that the world terms of trade line OT_w (in Figure 1), relevant when each union follows a policy of prohibitive tariff towards the rest of the world, is below the preunion situation terms of trade line OT_t. This means that the rest of the world countries will be oppositely affected by the formation of the customs unions: Some countries will have their terms of trade improved while others will have them deteriorated. Next, observe that the position of the multiple union situation terms of trade line $OT_{u.w}$ (Figure 2) under the assumption of nonprohibitive tariffs everywhere is below the preunion situation terms of trade line OT_t and between the free trade price lines[11] of the two unions (not shown).[12] There are two implications: First, even when the unions' trade with the rest of the world is not prohibitive, their combined effects on the latter (two) countries are opposite. This result is similar to the one just derived under the assumption of prohibitive union tariffs. However, it strengthens one of Arndt's main conclusions to which we called attention in the lead paragraph of this paper. Second, and more important, the above contrast is confined not only to countries of the rest of the world but also to countries within and between the unions themselves. For example, B_2, the passive member country of union u_1, will end up with a less favorable, while A_3, its counterpart in union u_2, *may*[13] end up with a more favorable terms of trade than the eventual world terms of trade. This implies that the passive member of a union need not necessarily lose in terms of trade. Rather the possibility exists that such a country, in spite of membership in a relatively high cost producing union, may end up with a gain. The interesting point is that opposite[14] experiences of gains and losses on the part of passive members of two different unions are obliterated in the single customs union analysis. No doubt the knowledge of such contrasts may be important to a country seeking membership in an emerging union in preference to one already in existence.

---

[11] These are the market clearing price lines.

[12] Located at the intersection of OA_2 with OB_2, and of OB_3 with OA_3.

[13] We stress here the more favorable terms of trade position of country A_3 to sharpen the contrast with country B_2. It should be noted that two other positions may exist, one of less favorable than and the other of equally favorable as the eventual world terms of trade. However, measuring by the area in the diagram applicable to more favorable, compared to less and equally

IV. Comparative Postunion (u_1 and u_2 vs. U) Situations

In the composite union (U) situation, the excess offer curve is derived from the free trade group offer curves OW''_A (of A_2 and A_3 countries) and OW''_B (of B_2 and B_3 countries). Assuming that this excess offer curve relates to the Y commodity originating from the A countries and that the tariff of the composite union is the same as that of the multiple unions towards the rest of the world, we have two versions of this excess offer curve, one designated as $OE_{U,f}$ based on free trade and the other $OE_{U,t}$ based on tariff calculations. Integrating $OE_{U,t}$ with OA'_1, the offer curve $OI''L''W_U$ is derived, some part ($I''L''$) of which is linear as usual. The world terms of trade are then obtained at the intersection of $OI''L''W_U$ and OB'_1, given by the slope of the line $OT_{U.w}$. The construction of Figure 2 under discussion shows that $OT_{U.w}$ and OT_t coincide, but this need not necessarily be the case.[15]

The interesting point to note however is the relative slope of the $OT_{U.w}$ and $OT_{u.w}$ lines: Figure 2 shows that the former line is steeper than the latter line. This becomes the basis of the conclusion that the excess Y producing country of a union in the $T_{u.w}$ situation has its terms of trade improved relative to the excess X producing country of the other union over the $T_{U.w}$ situation.[16] This of course assumes that in the latter situation, the union on the whole is excess Y producing. If not, i.e., if it is excess X producing,[17] the above conclusion will be reversed. In any case, opposite terms of trade effects on members of two different unions are once again revealed.

V. Welfare Implications[18]

We begin by comparing the $T_{u.w}$ with the T_t situation. Assume that the member countries of the union u_1 trade with each other at terms of trade given by the slope of the line Oa. Assume further that at these terms of trade, there is still some

---

favorable, terms of trade positions, the former position appears less probable.

[14] Of course the contrast would disappear had we to assume that both the unions generated excess offers in the same good. In that case, one of the member countries of each union would buy the above-indicated good from its free trade partner relatively dearer than the eventual world terms of trade would suggest. However, this case is a theoretical case for which practical justification may be difficult to find especially because of the similarity of the unions, and hence the scope for little or no bargaining between them.

[15] The important point to note in this connection is that there are also cases in which $OT_{U.w}$ may lie above as well as below OT_t; in the latter cases, it will still be above $OT_{u.w}$ under the assumed slope of the indifference curves, relative tariffs and costs. Thus, the ensuing argument about the comparative terms of trade effects in the multiple and the single union situations will still hold.

[16] It can be easily seen, in the light of the last footnote, that while the possibility exists for the rest of the world countries to have opposite terms of trade effects from the formation of a customs union or unions, two subpossibilities may arise. In one, the gaining country (viz., Y exporting A_1 country) increases the gain as the number of unions increases; in the other, the gaining country (viz., X exporting B_1 country) changes into the losing country with increase in the number of unions.

[17] And the two unions in the multiple union case are excess Y (viz., u_1) and excess X (viz., u_2) producing, as has been assumed above.

[18] This section is based on Figure 3 which is the same as Figure 2 except for some newly shown price lines.

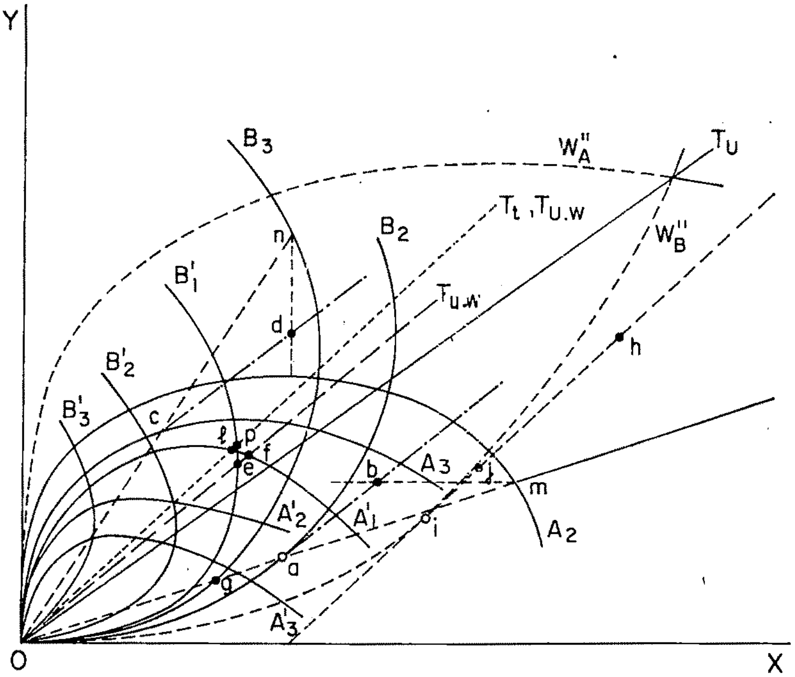


FIGURE 3

excess offer of Y left for which the possible market is outside this union and the price is the world price (given by the line) ab. Parallel to $OT_{u.w}$, this line has one extreme a which is the point of intersection of Oa with the offer curve OB_2, and the other extreme b which is the point of intersection of ab with bm. This last line is parallel to the X-axis and one of its extreme points m is the point at which Oa extended intersects the offer curve OA_2. Then bm will equal the total tariff expressed in units of X goods collected by this union. The final trading positions of country B_2 and A_2 of this union will clearly be at points a and b, respectively.

Similarly, assuming Oc to be the terms of trade line for countries of the other union and nd to be the total tariff expressed in units of Y goods and cd parallel to $OT_{u.w}$, the points c and d will represent the final trading positions of country A_3 and B_3 of this union. However, the positions of the rest of the world countries will be at points e and f, being the points of intersection of the respective country's offer curve with the world terms of trade line $OT_{u.w}$.

On the other hand, the trading positions of the six countries in the preunion situation will be at points where their offer curves cut the terms of trade line OT_t.

Under usual assumptions about the shapes of respective country's indifference curves (not shown on the graph), and comparing the postunion and the preunion trade positions of each country by their levels of indifference, the following conclusions emerge:

(i) One country of each of one union (viz., u_1) and the nonunionized "rest of the world" will be better off while the other country worse off,[19] both the gaining countries being the Y exporting A countries;

---

[19] This contrast (between the member countries of union u_1) will be sharpened the more we move to the market clearing intraunion prices; for movements away

(ii) Gains or losses of both the member countries of the other union will be rather ambiguous.

However, the ambiguity in (ii) above can be removed by pinpointing the relative slopes of the intraunion and postunion world terms of trade lines. Assuming that the former line lies below the latter, which will mean that the imports of country A_3 from its union partner are a bargain in the light of the eventual world prices, this country will gain and its partner lose; in a reverse case,[20] the opposite will be the effect.[21] The case in which the status quo will be maintained uniformly throughout the union is when the intraunion and world terms of trade lines coincide, the chance of which appears at best to be remote.

We now turn to a comparison of the postunion situations. One situation concerns the two unions u_1 and u_2, and the other the single union U. The comparison is supposed to shed some light on the direction in which the results of the usual single union analysis will be modified in the event of an additional union.

Under the single union approach, two sets of results somewhat interchanging and opposite in nature will follow depending on two kinds of assumptions made. We will briefly discuss one such assumption and the implications thereof, merely touching on the other assumption and its implications.

We assume that the composite union leads to excess[22] offers in Y. Starting with the intraunion terms of trade line *Ogaim* and drawing a line *ijh* parallel to $OT_{U.w}$, which goes through the point i at which *Ogaim* intersects the group offer curve OW''_B, we can obtain a point h on the line *ijh*. This point[23] will denote the final trading position of the A countries of the union, being equivalent to the point j for the A_2 *and* point i for the A_3 country. As the figure shows, a and g will indicate the trading positions of B_2 and B_3 countries. The trading positions of countries A_1 and B_1 will be determined at points where, as before, their offer curves are intersected by the line $OT_{U.w}$.

Using appropriate indifference curves corresponding to the final trading positions of various countries under the composite and multiple union situations, one can easily conclude the following as to their gains or losses in the latter over the former situation:

(a) Country A_2 and A_3 definitely lose
(b) Country B_2 remains unchanged (by construction) but B_3 most probably[24] gains
(c) Country A_1 gains but B_1 loses

If, on the other hand, we assume that the composite union leads to excess[25] offers

---

from the assumed prices (as shown by the line *Ogaim*) towards the X-axis, the contrast will still be valid, although a small possibility will arise for the B_2 country to remain unchanged in welfare.

[20] The reverse case is when the intraunion terms of trade line lies above the postunion world terms of trade line $OT_{u.w}$. It is important to qualify the result described in the text. This result will obtain when the intraunion terms of trade line is sufficiently above the line $OT_{u.w}$. If it is just above, possibilities exist for the welfare of both the member countries to remain either unchanged or even to decline!

[21] The possibilities of contrasting welfare effects (of the formation of the multiple customs unions) on members of the two unions should now be clear. For instance, for union u_1, the welfare of country A_2 improves, that of B_2 deteriorates; for union u_2, the welfare of country A_3 declines, that of B_3 improves.

[22] In order to make results of the composite and multiple unions reasonably comparable, we need an enabling assumption in the latter case, the sense of which is that the two unions, if left to trade between themselves only, will produce excess offers in commodity Y.

[23] Which is determined at the intersection of the line *ijh* and the line drawn parallel to the X-axis from the point at which the group offer curve OW''_A and the line *Ogaim*, both extended, intersect.

[24] There is a possibility that this country could remain unchanged in welfare.

[25] See fn. 22 for the enabling assumption suitable for this case.

in X, whereas the multiple union leads to excess offers in Y on the part of one union but to that in X on the part of the other, then starting with the intra-(composite) union terms of trade line *Ocn*, we can show that

(a′) Country B_2 and B_3 will definitely lose
(b′) Country A_3 will remain unchanged but A_2 will gain
(c′) Country A_1 will lose but B_1 will probably[26] gain

VI. Effects of Union Tariff Changes: Multiple Union Case

So far we have derived results assuming equal tariff rates for the two unions. The question will arise as to the nature of changes in the foregoing results when instead we assume unequal rates, all other things remaining the same as before. We discuss below both the terms of trade and welfare effects.

For terms of trade effects, refer to Figure 2; for welfare effects, see Figure 3. It will become immediately clear that the $OT_{u.w}$ line will be pulled up or down[27] depending upon union u_1 having the smaller or the larger tariff. But, in any case, the area over which it will move will not be too wide.[28]

Our first observation is that unequal tariff rates of any kind will not change the (qualitative) results for union u_1 derived under identical tariff rates: Country A_2, the country with the excess offer, will gain and country B_2, the passive country, will lose in both terms of trade and welfare. For union u_2, the passive country A_3 will have, as before, an improved, unchanged, or deteriorated welfare position no matter which union has the larger tariff. But, whereas in the case of the excess Y producing union with the larger tariff, the possibilities of terms of trade losses will increase, in the reverse case, all prior possibilities[29] will continue to hold, even though the magnitudes of losses or gains, as the case may be, will increase for this country. For its partner, however, much of whatever was found before to be the welfare effects will be qualitatively valid except for two modifications. One is when the excess Y producing union happens to be the one with the larger tariff, in which case the possibilities of losses will increase; the other is when the excess X producing union is with the larger tariff, in which case the possibilities of gains will increase. In fact, gains or losses will both be possible as in the case of identical tariffs but changes of one will increase in one set of circumstances, while those of the other increase in the other set of circumstances.

Our next observation is on the rest of the world countries. Here the same effects as in the case of equal union tariff rates will follow as long as union u_1 has the larger tariff: The contrasts between the (gaining) A and (losing) B countries in both the terms of trade and welfare effects will be maintained, even sharpened. But when the other union has the larger tariff, some ambiguities in both effects will appear, unlike the previous case. Now, in addition to possibilities that country A_1 gain and B_1 lose, there will be some new possibilities with reverse effects.

VII. Conclusion

We have derived in this paper some

---

[26] Country B_1 remaining unchanged in welfare is also a possibility.

[27] A similar downward tendency in the $OT_{u.w}$ line with corresponding terms of trade and welfare implications for the union and nonunion countries will follow in cases of differently constituted unions; for instance, unions u_3 and u_4 (see fn. 7) under equal union tariffs. Unfortunately, when such tariffs are unequal, the results will no longer be unambiguous. These cases, not shown on the graph, may easily be construed.

[28] Such movements will be limited by the location and curvature of the offer curves, the tariff rates, the excess offers and relative cost conditions.

[29] See particularly fn. 17 for such possibilities in the equal tariff case.

effects in the multiple union situation. These effects are basically of contrasting types in their relation to the rest of the world, and to intra- and interunion countries. Special attention has been paid to the interunion contrasts, naturally obliterated in the single union analysis. Knowledge of such contrasts, involving as they do terms of trade and welfare gains, seems to be quite important for countries seeking admission into already established unions or deliberating whether to form new unions.

Unfortunately, we could not tackle more general customs unions than those of two countries each. Nor could we tackle each member country of a union in the role of direct trade with the nonmember countries. (We have assigned this role, it may be recalled, exclusively to one of the two member countries of a union, the other member being assumed to trade wholly with its partner only.) All this would have been desirable, but the geometrical analysis, we are afraid, would have been too messy and complicated to warrant any simple and meaningful conclusion. Even with our simplified assumptions, the geometry looks very much involved at several places. In fine, our analysis provides some broad indications of possible results in certain multiple union situations only.

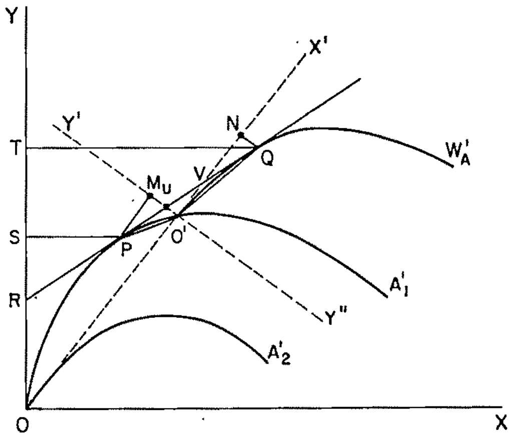


FIGURE 4

APPENDIX[30]

We now show a method of deriving group from individual offer curves modifying Arndt's construction.[31] It represents the final link in the integration of individual country offer curves. It serves two purposes: First, it shows (e.g., Figure 1) how the possible problem of perverse[32] tendency in reciprocal demand elasticity around the discontinuity point (e.g., *D*) in an Arndt-derived group offer curve can be overcome, and a traditional well-behaved elasticity[33] obtained in terms of the common tangent (e.g., *IL*) to the two wings (*OID* and DLW'_A) of the disjoint curve, providing for the linearity of the modified group offer curve. Second, it shows how the method of convex combination of trading processes (like *OI* and *OL*) will lead to optimality of results along that common tangent. Both these points are developed below. Some preliminary details are left out for which the reader is referred to Arndt. However, we retain as much of his notation as possible, even though at some cost of comparability with the present text.

---

[30] The subject matter of this Appendix was submitted in the form of a separate note entitled: "Customs Union and the Theory of Tariffs: Comment." For better exposition—thanks to expert editorial advice—that note is appended to the main paper, the present text.

[31] Arndt is not the first to have used this type of construction. At least two other authors before him have hinted at or actually constructed such aggregate reciprocal demand curves. See Elliot and Becker. We take here Arndt's work as the reference work because that is, to our knowledge, the most recent. We should hasten to add that the modification will not change any of Arndt's conclusions. For more details on this point, see fn. 4.

[32] Perverse in the sense that the possibility exists that one complete set of demand behavior characterized by greater than unitary, unitary, and less than unitary elasticity prior to that point could be repeated in that order immediately following that point. Such repetition will imply lack of rationality on the part of the subjects of aggregation.

[33] In the sense that the thus aggregated reciprocal demand curve will show only one complete set of demand behavior characterized by greater than unitary, unitary, and less than unitary elasticity taken in that order.

See Figure 4.[34] The two Y-exporting countries' offer curves are OA_1' and OA_2', each of the usual shape. For aggregation, we mark out the area OO' on country A_1 offer curve which lies above the terms of trade line OX' (being tangent to country A_2 offer curve). Obviously for all terms of trade steeper than the line OX', only country A_1 will be interested in trading Y for X, nothing coming from the other country. However, as Y becomes dearer than indicated by the slope of the line OX', the picture will change. For all such prices, A_2 will join A_1 and their total offers will swell along the new offer curve $O'QW'_A$ being obtained by the radial sum of the offers of A_2 along OA_2' and those of A_1 along $O'A_1'$. In summary, the group offer curve will work out to $OO'W'_A$ according to Arndt's construction.

Two important questions, as hinted at already, arise: First, what about the sharp discontinuity in the group offer curve at the point O' around which there may be some perverse[35] tendency in the values of the reciprocal demand elasticity? Second, could the integrated offer curve be elevated to an optimum path in the neighborhood of O', for, it is contended that Arndt's construction provides less than optimum results? Fortunately, both these questions can be tackled very simply by a single geometry as described below.

Shift the origin of the coordinate system from O to O'. At O', rotate the axes counterclockwise by the amount of the slope of the line OX'. The net effect will be to switch to the axes $OO'X'$ and $Y'O'Y''$ with origin at O' from axes OX and OY with origin at O. Draw a common tangent PQ to OPO' touching it at P and to $O'QW'_A$ touching it at Q such that it (the tangent) cuts $O'Y'$ and $O'X'$ at U and V, respectively, and when extended backward, cuts OY at R. From P, draw perpendiculars PS and PM on OY and $O'Y'$, respectively. From Q draw perpendiculars QT and QN on OY and OX', respectively.

It should be clear from the diagram that the transformation of the coordinate system has completely segregated the two wings of the disjoint group offer curve. In fact, we now have one wing OPO' belonging exclusively to the second quadrant, and the other wing $O'QW'_A$ belonging exclusively to the fourth quadrant of the new coordinate system, so that we can now view the former wing as a description of the conversion of X into Y, and the latter wing as that of Y into X, even though the original curve $OPO'QW'_A$ represents a unilateral conversion of X into Y. The important point is that both these wings basically represent the conversion processes of country A_1. But whereas it is an indigenous and in that sense a pure technology (of country A_1) underlying the conversion mechanism in the first wing, it is as if a hybrid and an improved technology has been used in the second wing. Country A_2 accounts for the hybrid technology and is indeed responsible for the improvement.

Following J. R. Hicks, pp. 80ff., and Nicholas Georgescu-Roegen, pp. 218–20, we can say that the common tangent PQ to the two wings will trace out the path of the optimum combination (integration) of the two kinds of trading processes[36] $O'P$ and $O'Q$. Also, the optimum profit from either conversion will be unique, given by $O'U$ expressed in Y units or $O'V$ expressed in X units. Consequently, PQ should replace those parts ($O'P$ and $O'Q$) of the previously determined group offer curve[37] which are clearly sub-optimal. Needless to add, the remaining parts (OP and QW'_A) will remain unaffected according to the method of the convex combination of processes. Thus, on the whole, is the group offer curve derived. It is $OPQW'_A$ and not $OPO'QW'_A$.

The preceding argument answers our second question. For the first question, the argument is as follows.

---

[34] Figure 4 has been extracted from Figure 1, focussing on the Y exporting countries offer curves only.

[35] Perverse in the sense of fn. 32.

[36] The Georgescu-Roegen, Hicks references deal with conversion through production, whereas here we consider conversion through trade. Analogy between the two problems arises from their underlying technical features: diminishing returns in their case and diminishing utility in ours.

[37] As we have described before, following Arndt's construction.

At P, the elasticity[38] is OS/OR; at Q, the elasticity is OT/OR, using in each case the original coordinate system. Since $OT>OS$, and $OS>OR$, elasticity at Q is higher than that at P, each being greater than unity. Further, elasticity of successive points beyond P will consistently increase up to the point Q after which it will decline (to the value unity and even less) along the curve QW'_A. On the other hand, along the curve OP, the elasticity will be consistently increasing with declining prices. So, then, along the whole curve $OPQW'_A$ elasticity of reciprocal demand will have a quite regular and traditional behavior.[39] This completes the answer to our first question.[40,41]

---

[38] Such a geometrical derivation of elasticity is well known. See, for example, James Meade.

[39] In the sense of fn. 33.

[40] Two questions will perhaps arise with regard to the propriety of the linear part irrespective of the question of its technical justifiability. One is related to the identity of the curve. On this, we submit that the aggregation produces some kind of admixture of the two countries just as an average does, leaving no singular identity of each country. The second question is related to the fact that the offer of the comparatively advantageous commodity for the other less advantageous commodity on the part of the countries aggregated is uniformly inflated. On this question, we submit that the aggregation yields a *potentially* efficient combination and the analysis should reflect such efficiency whether *or* not it is, in fact, realized in the international context.

[41] In a private communication, Arndt raised an issue with the dynamic justifiability of the aggregated offer curve. He said that the aggregation would involve a transition from a world economy of, say, two producers of a good to one integrated producer. Then, it would be a problem partly in oligopoly-monopoly theory and partly in decision making under conditions of centralized policy choices. If so, he felt that the perversity of elasticity as stated in our footnote 32, which he took to be only transitional, would disappear as time intervened. True. But our (modified) group offer curve is consistent with his well thought-out presumptions.

REFERENCES

S. W. Arndt, "Customs Union and the Theory of Tariffs," *Amer. Econ. Rev.*, Mar. 1969, *59*, 108–17.

G. S. Becker, "A Note on Multi-Country Trade," *Amer. Econ. Rev.*, Sept. 1952, *42*, 558–68.

C. A. Cooper and B. F. Massell, "A New Look at Customs Union Theory," *Econ. J.*, Dec. 1965, *75*, 742–47.

G. A. Elliot, "The Theory of International Values," *J. Polit. Econ.*, Feb. 1950, *58*, 16–29.

N. Georgescu-Roegen, "Leontief's System in the Light of Recent Results," *Rev. Econ. Statist.*, Aug. 1950, *32*, 214–22.

F. Graham, "The Theory of International Values Re-examined," *Quart. J. Econ.*, Nov. 1923, *38*, 54–86.

J. R. Hicks, *Value and Capital*, Oxford 1965.

R. G. Lipsey, "The Theory of Customs Union: A General Survey," *Econ. J.*, Sept. 1960, *70*, 496–513.

J. E. Meade, *A Geometry of International Trade*, London 1952.

M. Michaely, "On Customs Union and the Gains from Trade," *Econ. J.*, Sept. 1965, *75*, 577–83.

J. Vanek, *International Trade: Theory and Economic Policy*, Homewood 1962.

J. Viner, *The Customs Union Issue*, New York 1950.

The Dynamics of Inflation in Latin America, 1950–1969

By ROBERT C. VOGEL*

A decade after its publication, Arnold Harberger's study of Chilean inflation remains the best known empirical analysis of the causes of inflation in any Latin American country. Subsequent studies of Latin American inflation include only two major econometric analyses of its causes, and both are extensions of Harberger's Chilean model to inflation in Argentina (see Carlos Diaz-Alejandro and Adolfo Diz). Although such studies cannot completely resolve the controversy between the monetarists and structuralists (which also involves the relationship between inflation and growth, and opposing policy recommendations to deal with inflation), determining the causes of Latin American inflation is an important first step. By extending Harberger's model to sixteen Latin American countries for the period 1950–69, this paper investigates the impact of monetary factors on inflation and, in particular, the extent to which countries with widely varying inflationary and monetary experiences exhibit homogeneous behavior. In addition to reporting individual regressions for the different countries, the paper makes use of pooled regressions combining time-series and cross-section observations. Pooling data over several countries not only offers more degrees of freedom for hypothesis testing, but also provides a direct method for testing the applicability of Harberger's model throughout Latin America.

Contrary to popular opinion, all Latin American countries have not been experiencing rampant inflation. As shown in Table 1, seven of the sixteen countries included in this study have had rates of inflation averaging less than 4 percent per year during the period 1950–69. Of the remaining nine countries, four have experienced moderate inflation, averaging 5 to 13 percent per year, and only five have undergone substantial inflation, averaging more than 25 percent per year. The average rates of money supply growth also vary widely among countries, roughly paralleling the rate of inflation for each country.[1] Table 1 also reveals that the different countries have widely varying average velocities, but that the standard deviation in velocity for each country is relatively small. Differing average rates of inflation, money supply growth and levels of velocity suggest that Latin American countries have significant differences in institutions and patterns of behavior. In the face of such diversity, it is particularly important to investigate the extent to

---

* Senior staff economist, Council of Economic Advisers, on leave from Southern Illinois University. This paper was written while I was at Wesleyan University, and I am particularly indebted to James Hanson and Michael Lovell for their suggestions and encouragement. I would also like to thank Francisco Chaves and the Wesleyan Computer Center for their assistance with the calculations. Helpful comments on this paper were also made by Arnold Collery, William Gibson, G. S. Maddala, the managing editor, and an anonymous reviewer.

[1] Countries with higher rates of inflation and money supply growth also tend to have higher standard deviations for these two variables. Harry Johnson, among others, has suggested that variability in the rate of inflation, more than inflation itself, tends to hinder economic growth. However, Table 1 gives only limited evidence of an inverse relationship between either inflation or its standard deviation and the average rate of growth in real income for each country, and Johnson himself rejects such casual empiricism to resolve the monetarist-structuralist controversy.

TABLE 1—INFLATION AND RELATED STATISTICS FOR SIXTEEN LATIN AMERICAN COUNTRIES, 1950–69

	Rate of Inflation[a]		Money Supply Growth[a]		Real Income Growth[a]		Velocity	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Uruguay	43.0	34.9	40.1	33.3	0.7	6.3	5.96	0.82
Bolivia	41.3	55.9	41.5	58.1	3.0	22.8	11.77	5.58
Brazil	35.1	21.5	38.2	22.1	3.9	4.3	4.17	0.39
Chile	28.2	14.6	35.2	12.8	4.6	6.4	11.35	1.74
Argentina	26.4	23.4	24.6	11.8	2.4	7.3	4.14	1.01
Paraguay	12.5	17.5	15.4	14.2	5.5	4.1	11.27	1.21
Colombia	9.2	8.2	16.5	5.5	5.4	3.9	6.56	0.50
Peru	8.5	5.8	13.4	6.0	5.7	7.4	8.38	0.64
Mexico	5.3	5.3	11.3	4.3	6.9	4.6	8.22	0.52
Nicaragua	3.4	6.1	8.6	11.1	3.7	7.5	8.85	0.94
Ecuador	3.0	3.3	8.8	8.2	4.7	3.7	8.00	0.62
Honduras	2.1	3.4	8.0	8.7	4.0	3.9	10.21	1.11
Costa Rica	1.9	2.2	9.0	7.9	5.7	3.4	6.16	0.41
Guatemala	1.1	2.8	5.9	7.4	3.9	3.1	9.47	0.78
Venezuela	1.1	2.8	7.9	8.8	6.8	4.8	6.70	0.50
El Salvador	0.3	1.5	3.5	6.9	4.6	3.5	7.58	0.62

[a] Percent per year.

Note: S.D. is the standard deviation. Inflation is measured by the consumer price index; money supply is currency plus demand deposits; real income is nominal *GNP* deflated by the consumer price index.

which these countries respond homogeneously to the monetary factors thought to cause inflation.

The first section of this paper examines Harberger's model of Chilean inflation and its subsequent applications to Argentine inflation. The second section reports the results of the pooled and individual country regressions extending Harberger's model to the rest of Latin America. The third section reports the results of pooled regressions using alternative lag structures. The final section summarizes the principal findings and indicates their implications for the monetarist-structuralist controversy over inflation in Latin America.

I. Harberger's Model of Chilean Inflation and Its Extensions to Argentina

Given the long and varied history of Chilean inflation, it is not surprising that a number of economists have studied this subject (see, for example, Tom Davis and Albert Hirschman). However, only Harberger has developed an econometric model of the causes of Chilean inflation which can readily be generalized and estimated for all Latin America. As the dependent variable in this model Harberger chooses the rate of inflation (which he labels P_t) rather than the price level, after arguing that the rate of inflation is more important for policy decisions and is more difficult to explain (i.e., is less correlated with possible explanatory variables). Following a monetarist approach, Harberger introduces as independent variables the percentage change in the money supply during the current and preceding years (which he labels M_t and M_{t-1}) and the percentage change in real income during the current year (which he labels Y_t).[2] To allow for changes in velocity caused by changes in the expected costs of holding money, he introduces past changes in the rate of inflation ($P_{t-1}-P_{t-2}$, which he labels A_t) as a proxy for these expected changes.[3] To allow for the influence of structuralist factors, he includes a wage variable, the percentage change in the *sueldo vital* (the minimum salary for government white-collar employees, a major factor in wage agreements throughout the Chilean economy).

---

[2] Harberger chooses this particular structure for lagged adjustment to money supply changes because the coefficients of M_t and M_{t-1} sum to approximately one, and the coefficient of M_{t-2} is not statistically significant; but, as Diz indicates, such criteria may not be entirely appropriate or conclusive.

Harberger's study covers the period 1939–58 for the Chilean economy, and for the observations in his main regressions he uses annual percentage changes of the consumer price index, currency plus demand deposits, and real national income.[4] Ordinary least squares regression yields:

$$
\begin{aligned}
(1)\quad P_t = &\underset{(9.23)}{-\,0.32} + \underset{(.16)}{0.74M_t} + \underset{(.15)}{0.34M_{t-1}} \\
&\underset{(.31)}{-\,0.91Y_t} + \underset{(.12)}{0.20A_t} \qquad R^2 = 0.87
\end{aligned}
$$

(Standard errors are given in parentheses below each coefficient.) The sign, magnitude, and significance of each coefficient agree with Harberger's expectations, although A_t is significant only at the 5 percent level. When the wage variable is added, its coefficient is less than its standard error, and it does not appreciably affect the magnitudes or standard errors of the other coefficients, except that A_t is no longer significant at the 5 percent level. In discussing these findings, Harberger does not adopt the entirely monetarist position which his results would seem to support, but discussion of implications for the monetarist-structuralist controversy will be deferred to the final section of this paper.

As previously mentioned, the studies by Diaz-Alejandro and Diz are the only other major econometric analyses of the causes of inflation in a Latin American country, and both are extensions of Harberger's model to Argentina. Diaz-Alejandro introduces four independent variables into his regressions: money supply, real gross domestic product plus merchandise imports, hourly money wage rates in industry, and the peso/dollar exchange rate. He uses five different dependent variables (the wholesale price index, three components of that index, and the cost of living index for Buenos Aires), and all variables are expressed as annual rates of change, with observations for the years 1945–62.[5] Diaz-Alejandro's regressions have R^2s ranging from 0.77 to 0.92, and the coefficients generally are more than twice their standard errors and have the expected signs and magnitudes. The money supply coefficients are not significantly different from plus one, nor the national product coefficients from minus one; while changes in both wages and the exchange rate have a positive, though less than proportionate, impact on the rate of inflation.[6]

Diz, like Diaz-Alejandro, cites Harberger's work on Chile as the basis for his study of the causes of inflation in Argentina. Diz's regressions cover the period 1935–62 and use quarterly rather than annual observations, with all variables expressed as first differences of logarithms. The independent variables in his regressions are: money supply (with two definitions, one including and the other excluding savings and time deposits), real income, an index of nominal wages, the official peso/dollar exchange rate, and a measure of price expectations.[7] The dependent variables are the wholesale price and cost of living indices. Diz's various regressions all yield similar results, with R^2s ranging from 0.60 to 0.76. Current money supply coefficients, as well as some lagged money supply coefficients, are highly significant and indicate initial overshooting and adjustment continuing more than two years.[8] The income coefficients are also highly significant, ranging from -0.445 to -0.845, while the price expectations coefficients are positive and significant. The wage coefficients are not significant and the exchange rate coefficients, though significant, imply a highly inelastic response of prices to changes in the exchange rate. Like Harberger, both Diaz-Alejandro and Diz relate their findings to the monetarist-structuralist controversy; but, as previously indicated, discussion of this controversy will be deferred to the concluding section.

---

[3] As will be indicated, independent variables containing lagged price changes can be given alternative interpretations (see Harberger, p. 241 n.). Interest rates are not introduced as explanatory variables because they do not accurately reflect the costs of holding money; in Chile, as in other Latin American countries, interest rates are subject to extensive and varying government controls.

[4] In other regressions Harberger uses the wholesale price index and various components of the wholesale and consumer indices as the dependent variable, and he uses both quarterly and annual data. Because such a variety of data is not available for all Latin American countries, the regressions in the present study are limited to the consumer price index and to an annual basis. Moreover, Harberger's other regressions generally yield results similar to the regression reported above.

[5] In contrast to Harberger, no lagged rates of change are included; all such explanatory variables were found by Diaz-Alejandro to be insignificant, as were other possible explanatory variables.

[6] In contrast to Harberger, the intercepts in Diaz-Alejandro's regressions are unexpectedly significant and range from −10.5 to −20.2. This suggests that, when other things are equal, the price level will decline by 10–20 percent per year due to declining velocity; and this can be compared to similar findings in the present study.

II. Regression Results Using Harberger's Model

The present study extends Harberger's model to sixteen Latin American countries for the period 1950–69.[9] To facilitate interpretation of the regression results, it is useful to show how this model is related to the quantity theory of money. Beginning from the identity:

$$MV \equiv PY \tag{2}$$

where M is the exogenously determined money supply, V is velocity, P is the price level, and Y is real income, and assuming a simple money demand function:

$$V^{-1} = Y^aC^b \tag{3}$$

where the ratio of real income held as real balances, the inverse of velocity, depends on the level of real income and on the expected costs of holding real balances C, we obtain:

$$M = PY^{1+a}C^b \tag{4}$$

Solving for P and transforming to percentage rates of change (denoted by a prime) yields:[10]

$$P' = M' - (1+a)Y' - bC' \tag{5}$$

This can be compared with the equation actually estimated in both Harberger and this study:

$$P'_t = k + dM'_t + eM'_{t-1} - (1+a)Y'_t + b(P'_{t-1} - P'_{t-2}) \tag{6}$$

The absence of a term in front of M' in equation (5) results from assuming that the demand for real money balances depends only on real variables (i.e., there is no money illusion) and that adjustment is instantaneous. In equation (6) dM'_t and eM'_{t-1} are introduced to capture the lagged adjustment of prices to changes in the money supply. The term $(1+a)$ indicates that the estimated coefficient for real income should deviate from minus one only to the extent that the level of real income affects the ratio of real balances held to real income. In Harberger and the present study, past changes in the rate of inflation is the main variable affecting the expected costs of holding real balances, and hence velocity, so that C in equation (5) is replaced by $P'_{t-1}-P'_{t-2}$ in equation (6). Trends in velocity resulting from trends in the expected costs of holding real balances are captured by the intercept term, k.

---

[7] See Diz, pp. 114–15, for a description of his method for constructing a measure of price expectations from past prices. Like Harberger, Diz mentions that the price expectations variable may also be a proxy for additional lags of the money supply, but he neglects to mention other possible interpretations and the statistical problems which arise from lagged dependent variables, even if his method is deemed appropriate.

[8] Diz states that prices may overshoot in their response to money supply changes and argues, in contrast to Harberger, that it is not necessarily appropriate to stop adding lags of the money supply when the coefficients sum to one or when an additional lagged coefficient is not significant.

[9] The countries included are those listed in Table 1. Other Latin American countries have been omitted: the Caribbean island nations, British Honduras and the Guyanas because of political dependence and lack of data, and Panama because of the lack of an independent money supply.

[10] Harberger, p. 220, mentions that the demand for money or velocity, rather than prices, might have been chosen as the dependent variable. The possibility that these variables are jointly determined suggests the existence of the well-known simultaneous equations problem. However, following Harberger, Diz, and Diaz-Alejandro, the present study has chosen to neglect this problem and to use ordinary least squares as a first approximation in explaining Latin American inflation.

The regressions in the present study are based primarily on data from *International Financial Statistics*. Inflation is measured by the consumer price index, and the money supply is defined as currency plus demand deposits.[11] *The Yearbook of National Accounts Statistics* is the main source for data on real income (nominal *GNP* in the currency of each country, deflated by that country's consumer price index). All variables are expressed as percentage changes.[12] Structuralist variables, such as changes in wages and exchange rates, have not been included because comparable data do not exist for a sufficient number of Latin American countries.[13] In addition, Harberger's model, like its adaptations by Diz and Diaz-Alejandro, is essentially monetarist, and the added structuralist variables are not an integral part of the model and do not contribute appreciably to explaining inflation in either Chile or Argentina. Moreover, the main purpose of the present study is to investigate the extent to which a purely monetarist model, and in particular changes in the money supply, can explain inflation throughout Latin America.

Table 2 presents the results of the pooled regressions for the sixteen Latin American countries.[14] Intercept terms are treated in four different ways, and each alternative implies a different pattern of behavior for velocity: with a common intercept, the trend in velocity is the same for all countries and years; with year dummies or country dummies, each year or country respectively has a different trend in velocity; and with a zero intercept, velocity is constant. Except for the income coefficients, varying the treatment of the intercept term has little effect on the results. The addition of a common intercept term or dummy variables contributes little to explanatory power, as indicated by the uniformly high $\bar{R}^2$s. Moreover, considering the diversity among Latin American countries, these $\bar{R}^2$s are remarkably high—almost as high as the R^2s obtained by Harberger for Chile alone.

In agreement with Harberger's findings, the coefficients of both current and lagged

---

[11] Regressions were also run with the money supply defined as currency and then as currency plus demand deposits plus quasi money. These regressions, which gave similar results to those reported here, are available from the author on request.

[12] Percentage changes are used in the regressions in the following tables to facilitate comparison with Harberger's results. All regressions in the present study have also been run with first differences of logarithms (see Diz) which is an essentially equivalent transformation and yields the same results. Furthermore, regressions using rates of change give well-behaved error terms, in contrast to regressions using levels which show high positive serial correlation.

[13] Even when data for structuralist variables do exist, they are generally not comparable among countries; exchange rates are associated with varying exchange controls tariffs, etc. (even in the absence of multiple exchange rates), and wage rates have varying definitions and degrees of government intervention.

[14] Data on 16 countries over 20 years should yield 320 observations, but 28 observations are lost due to incomplete data and 48 more due to calculating percentage changes and introducing lagged variables. Thus, only 244 observations are available for the pooled regressions. Table 3 indicates the number of observations for each country.

TABLE 2—HARBERGER'S MODEL FOR CHILEAN INFLATION EXTENDED TO ALL LATIN AMERICA: POOLED REGRESSIONS FOR SIXTEEN COUNTRIES, 1950–69

	Intercept	M'_t	M'_{t-1}	Y'_t	$P'_{t-1}-P'_{t-2}$	$\bar{R}^2$	$\bar{S}_e$	D.W.
Pooled with common intercept	−.031	.586** (17.0)	.407** (11.1)	−.298** (3.1)	.014 (0.4)	.82	.104	1.55 1.88
Pooled with year dummies	−.038	.595** (16.4)	.418** (10.8)	−.238* (2.4)	.032 (0.8)	.82	.103	1.59 1.98
Pooled with country dummies	−.019	.549** (15.0)	.377** (9.5)	−.280** (2.9)	.018 (0.5)	.82	.103	1.60 1.91
Pooled with zero intercept	—	.561** (16.5)	.365** (10.6)	−.456** (5.7)	.015 (0.4)	.81	.105	1.46 1.90

Note: M is currency plus demand deposits; Y is real *GNP*; and P is the consumer price index. Primes indicate annual percentage changes, and subscripts indicate current or lagged variables. Intercepts for pooled regressions with dummy variables are averages of the intercepts for individual years or countries. The *t*-values are given in parentheses below each coefficient; those marked with one asterisk (*) are significant at the 10 percent level, and those marked with two asterisks (**) at the 1 percent level. $\bar{R}^2$ and $\bar{S}_e$ are respectively the multiple correlation coefficient and the standard error of the estimate, both adjusted for degrees of freedom. For the pooled regressions the first Durbin-Watson statistic (*D.W.*) corresponds to data ordered by countries (and by years within each country), and the second to data ordered by years (and by countries within each year). Pooled regressions have 244 observations, and n indicates the number of observations for individual country regressions.

money supply are highly significant, and their magnitudes indicate that more of the adjustment of inflation to changes in the money supply occurs during the first year and that adjustment appears to be virtually complete by the end of the second year. Also like Harberger's findings, the coefficients of lagged changes in the rate of inflation are positive but not significant, suggesting either that expected changes in the rate of inflation do not appreciably affect the costs of holding real balances or that past changes in the rate of inflation are not a good proxy for expected changes. The income coefficients are negative and significant, but their magnitudes never approach the value found by Harberger, and this seems to imply that the demand for real balances is inelastic with respect to real income. Simultaneously, the negative intercept terms indicate a downward trend in velocity, which seems to imply that the costs of holding real balances are falling over time. However, constraining the intercept to be zero increases the magnitude of the income coefficient, and this apparent tradeoff suggests that falling velocity may be a proxy for rising real income.[15] The use of Durbin-Watson statistics in the context of pooled regressions is rather unusual and will be discussed in the next section.

Table 3 presents the results of the individual country regressions, with countries listed by rates of inflation, from highest to lowest. Although these results generally agree with the findings from Table 2, there are clearly some differences among countries. The $\bar{R}^2$s range from .00 to .87, but more than half are above .50. Not all the coefficients of money and income are significant, but virtually all have the expected sign. The coefficients of lagged changes in the rate of inflation are varied

---

[15] Falling velocity may be a better proxy for rising real income than is observed income because of the difficulties in measuring income accurately, and especially because permanent (i.e., expected) income may be a more appropriate concept than current observed income. In the next section, lagged income variables will be introduced to explore, at least partially, the appropriateness of permanent income.

TABLE 3—HARBERGER'S MODEL FOR CHILEAN INFLATION EXTENDED TO ALL LATIN AMERICA: INDIVIDUAL REGRESSIONS FOR SIXTEEN COUNTRIES, 1950–69

	Intercept	M'_t	M'_{t-1}	Y'_t	$P'_{t-1}-P'_{t-2}$	$\bar{R}^2$	$\bar{S}_e$	D.W.	n
Uruguay	−.026	.339 (1.5)	.862** (5.7)	−1.499 (1.7)	.078 (0.1)	.86	.131	1.53	11
Bolivia	−.057	.664** (6.7)	.397** (3.2)	.266 (0.9)	.089 (0.8)	.87	.204	1.60	17
Brazil	.014	.520* (3.0)	.411* (2.8)	−.416 (0.6)	.168 (0.7)	.79	.097	2.27	16
Chile	.113	.494 (1.2)	−.000 (0.0)	−.478 (0.6)	.241 (0.9)	.03	.127	2.00	11
Argentina	.038	.842* (2.2)	.304 (0.8)	−2.357** (3.6)	−.041 (0.3)	.56	.163	1.66	17
Paraguay	.008	.287* (1.9)	.361* (2.7)	−.438 (0.9)	−.139 (1.3)	.62	.049	1.10	15
Colombia	.132	.129 (0.5)	.234 (0.9)	−1.780** (5.0)	.107 (1.0)	.59	.052	2.27	16
Peru	.044	.437** (3.4)	.236* (1.9)	−.803** (7.3)	.099 (1.2)	.84	.025	1.56	15
Mexico	−.006	.458 (1.2)	.410 (1.2)	−.737 (1.4)	−.232 (1.0)	.00	.048	1.49	16
Nicaragua	.033	.071 (0.9)	.115 (1.5)	−.562** (5.0)	−.065 (0.7)	.71	.027	2.51	17
Ecuador	.001	.220** (3.4)	.338** (5.4)	−.528** (3.4)	.195 (1.4)	.68	.015	1.61	17
Honduras	.018	−.008 (0.1)	.212* (2.9)	−.268 (1.4)	.122 (0.8)	.42	.022	2.30	17
Costa Rica	.024	.023 (0.6)	.092* (2.5)	−.244* (1.8)	.076 (0.7)	.33	.012	1.75	17
Guatemala	.036	.137 (1.6)	−.060 (0.8)	−.685** (3.5)	.346 (1.3)	.34	.023	1.48	17
Venezuela	.021	−.032 (0.4)	.144* (2.9)	−.313* (2.2)	.197* (1.8)	.63	.015	2.26	16
El Salvador	.011	.059 (0.6)	.111 (1.3)	−.222 (0.9)	−.075 (0.2)	.00	.016	2.35	9

Note: See Table 2.

in sign and, like Table 2, are rarely significant. The intercept terms, indicating the trend in velocity for each country, range from −.057 to .132, but more than half lie close to zero, between −.03 and .03. There appears to be no relationship between a country's rate of inflation and either its intercept term or its coefficient for lagged changes in the rate of inflation. However, there is a slight tendency for the coefficients of money to be larger and more significant for high inflation countries, and for the coefficients of income to be more significant for low inflation countries. The Durbin-Watson statistics range from 1.10 to 2.51, but more than half lie close to two, between 1.60 and 2.40. However, with a maximum of only 17 observations, it is almost impossible either to accept or to reject the hypothesis of serially correlated residuals. Finally, it is worth noting that the performance of the Harberger model is worse for Chile than for almost any other Latin American country.

The F-tests were run to determine the extent of homogeneous behavior among different countries and years; that is, whether the introduction of separate intercepts (dummy variables) or differing slopes (individual regressions) can significantly reduce the residual sums of squares (see Gregory Chow and Edwin Kuh).[16] In the pooled regressions in Table 2 the common intercept term is significantly different from zero at the 1 percent level, but introducing separate intercept terms for individual countries or years does not contribute significantly to explaining rates of inflation, even at the 10 percent level. This implies a significant common downward trend in velocity for all Latin American countries, a trend which does not vary significantly despite the diversity among these countries. On the other hand, introducing different slope coefficients for each country or year does reduce the residual sums of squares significantly, but barely at the 1 percent level. This heterogeneity in the effect of money and income on inflation has already been suggested by the differences among countries shown in Table 3, but the $\bar{R}^2$ obtained by allowing all coefficients to differ among countries is only .03 higher than the $\bar{R}^2$s obtained in the pooled regressions in Table 2. Differences among coefficients for individual years are more difficult to interpret, since the regressions for individual years (not reported here) display no discernible patterns, but simply random variations.

III. Regression Results Using Alternative Lag Structures

The previous discussion and findings suggest several important reasons for testing lag structures different from those in Harberger's model. The use of lagged changes in the rate of inflation as an independent variable in Harberger's model introduces the well-known statistical difficulties associated with lagged dependent variables.[17] Moreover, when Harberger's model is extended to all Latin America, the coefficients of current income in Table 2 are unexpectedly low, and introducing lagged income variables may indicate the appropriateness of permanent income. Finally, an additional lag of the money supply is introduced to test Harberger's finding that the rate of inflation is influenced by the money supply of only the current year and the preceding year.

The first four regressions in Table 4 use percentage changes in the money supply and real income for the current and two preceding years as the independent variables to explain the rate of inflation, and omit any variables for lagged price changes. The four different treatments of the intercept term are the same as in Table 2 and yield results quite similar to the comparable regressions in that table, although in Table 4 the $\bar{R}^2$s are slightly higher and the intercepts slightly more negative. The money supply coefficients are also quite similar to those in Table 2, although the coefficients of current money supply in Table 4 are slightly smaller and the coefficients of money supply lagged one year slightly larger.[18] As Harberger found for Chile, the coefficients of money supply lagged two years are not significant. The coefficients of current income in Table 4 are almost identical to those in Table 2, and the coefficients of income lagged two years are not significant.

The most important difference from Table 2 is that in Table 4 the coefficients of income lagged one year are positive and highly significant. Because lagged income variables were introduced to test whether permanent income is superior to current income in explaining inflation, the coefficients of lagged income (like current income) were expected to be negative. How can the opposite finding be rationalized? It

---

[16] Details of these F-tests are available from the author on request.

[17] These difficulties have most often been discussed in the context of estimating distributed lags; see, for example, Zvi Griliches or Wayne Fuller and James Martin.

[18] This difference may be due to the fact that lagged price changes in Table 2 tend to act as a proxy for lags in the money supply (see Harberger, p. 241 n.).

TABLE 4—ALTERNATIVE LAG STRUCTURES TO EXPLAIN LATIN AMERICAN INFLATION:
POOLED REGRESSIONS FOR SIXTEEN COUNTRIES, 1950–69

	Intercept	M'_t	M'_{t-1}	M'_{t-2}	Y'_t	Y'_{t-1}	Y'_{t-2}	P'_{t-1}	P'_{t-2}	$\overline{R}^2$	$\overline{S}_e$	D.W.
Pooled with common intercept	−.051	.538** (14.8)	.434** (10.8)	.036 (0.9)	−.289** (3.2)	.323** (3.4)	.026 (0.3)	—	—	.82	.102	1.49 1.83
Pooled with year dummies	−.062	.533** (13.9)	.461** (10.7)	.041 (1.0)	−.234* (2.5)	.386** (4.0)	.037 (0.4)	—	—	.83	.100	1.48 1.97
Pooled with country dummies	−.044	.479** (12.7)	.404** (10.0)	.036 (0.9)	−.261** (2.9)	.439** (4.7)	.117 (1.4)	—	—	.84	.098	1.55 1.78
Pooled with zero intercept	—	.536** (14.3)	.406** (9.9)	−.026 (0.7)	−.466** (5.6)	.152* (1.3)	−.089 (1.1)	—	—	.81	.105	1.42 1.91
Pooled with common intercept	−.038	.430** (10.8)	.261** (5.3)	−.147** (2.8)	−.306** (3.3)	.561** (5.7)	−.004 (0.0)	.349** (4.8)	.115* (1.9)	.84	.096	1.94 1.91
Pooled with year dummies	−.046	.428** (10.2)	.292** (5.6)	−.130* (2.4)	−.229* (2.4)	.617** (6.0)	−.065 (0.6)	.372** (4.9)	.067 (1.1)	.85	.095	1.95 2.01
Pooled with country dummies	−.035	.396** (9.7)	.265** (5.3)	−.120* (2.3)	−.287** (3.1)	.617** (6.2)	.103 (1.0)	.278** (3.7)	.119* (2.0)	.85	.095	1.94 1.86
Pooled with zero intercept	—	.417** (10.3)	.222** (4.6)	−.209** (4.3)	−.432** (5.2)	.465** (4.8)	−.096 (0.9)	.390** (5.4)	.123* (2.0)	.84	.098	1.94 1.97

Note: See Table 2.

may be that in Latin America permanent (i.e., expected) income is a function of the change in income ($Y_t - Y_{t-1}$), rather than lagged income with declining weights over time. Moreover, lagged income variables are only a rough proxy for permanent income; and the persistence of negative intercepts, which imply a downward trend in velocity, may in reality indicate an upward trend in permanent income which is not adequately represented by lagged income variables.

In the remaining four regressions in Table 4, lagged price changes (P'_{t-1} and P'_{t-2}) have been added to the independent variables. For the four treatments of the intercept term, the intercepts are somewhat less negative and the $\overline{R}^2$s somewhat higher than in the first half of the table. The increase in $\overline{R}^2$s is mainly attributable to the highly significant coefficients of price changes lagged one year, while the coefficients of price changes lagged two years are much smaller and less significant. The addition of lagged price variables increases somewhat the coefficients of income lagged one year but does not appreciably affect the other income coefficients. The coefficients of money supply are all somewhat reduced, but particularly the coefficients of lagged money supply (for example, M'_{t-2} becomes negative and significant).

The importance of lagged price variables can be attributed to one or more of the following factors: 1) lagged adjustment in the rate of inflation to changes in money supply and income; 2) changes in the rate of inflation due to expectations based on past changes in prices; and 3) errors in the specification of the model, particularly the exclusion of significant variables (for example, structuralist causes of inflation). Both Harberger and Diz interpret their lagged price variables in the context of an expectations model, but the difficulties in distinguishing between lagged adjustment and expectations models are well known (see, for example, Roger Waud). Moreover, errors in specification cannot be ruled out because statistical problems arise when lagged dependent variables are introduced; in particular, serial correlation of residuals due to the exclusion of significant variables may be picked up by lagged price variables.

No definite conclusions can be drawn about serial correlation of residuals from the Durbin-Watson statistics in Table 3 because the individual country regressions do not provide enough observations. Therefore Durbin-Watson statistics, though normally used only in the context of time-series regressions, have been calculated for the pooled regressions in Tables 2 and 4 by ordering the observations in two different ways: by countries according to rates of inflation (and by years within each country); and by years (and by countries according to rates of inflation within each year). For each pooled regression the first Durbin-Watson statistic tests for differences in residuals among countries and serial correlation of residuals within countries, while the second tests for differences in residuals among years and nonrandom residuals related to rates of inflation. The second Durbin-Watson statistics are sufficiently close to two to suggest that there are no significant differences in residuals among years and that residuals are randomly distributed with respect to rates of inflation.[19] The first Durbin-Watson statistics are also close to two when lagged price variables are included,[20] but sufficiently far from two when lagged price variables are excluded (or in the Harberger model) to suggest significant serial correlation of residuals within countries.[21]

These findings support the hypothesis that the significance of lagged prices is due to errors in specification, but the two alternative hypotheses cannot be completely ruled out, since serial correlation and the significance of lagged prices may be due to the importance of price expectations or lagged adjustment, rather than to the exclusion of other significant variables. Furthermore, solving the last four equations in Table 4 for the implied lags yields essentially the same patterns over time as the comparable regressions without lagged prices. Thus, regardless of the reason for their significance, adding lagged price variables does not appreciably change the pattern of response over time of inflation to changes in the money supply and real income.

The F-tests were also run for the regressions using alternative lag structures to determine the extent of homogeneous behavior among different countries and years. The F-tests for these regressions yield essentially the same results as the F-tests for the regressions using the Harberger model. The common intercept terms are significantly different from zero at the 1 percent level; separate intercepts for individual countries or years are not significant at the 10 percent level; and slope coefficients are significantly different for individual countries and years at the 1 percent level. In general, the F-ratios are somewhat higher than for the Harberger model, indicating greater heterogeneity for the alternative lag structures, especially among slope coefficients in the individual regressions. However, the main findings remain the same: a significant downward trend in velocity which does not vary significantly among countries, and no discernible patterns in the heterogeneous slope coefficients for individual countries and years.[22]

---

[19] Because significance levels have not been established for Durbin-Watson statistics in the present context of pooled regressions, these conclusions are based on significance tables for Durbin-Watson statistics in the normal context of time-series regressions.

[20] Marc Nerlove and Kenneth Wallis have pointed out that the Durbin-Watson statistic will be biased toward two when lagged dependent variables are included.

[21] Differences in residuals among countries, which could also cause the first Durbin-Watson statistic to differ from two, have been removed in the pooled regressions with country dummy variables.

[22] The problem of heteroscedasticity should be discussed briefly. As indicated in Table 1, Latin American

IV. Conclusions

Despite the extreme diversity among Latin American countries in rates of inflation, money supply growth and velocity, both the Harberger model and the alternative lag structures yield remarkably good results. For all the pooled regressions, the $\bar{R}^2$s indicate that 80 to 85 percent of the variance in the rate of inflation is explained; and, although the F-tests reveal some heterogeneity among countries and years, the $\bar{R}^2$s are not appreciably increased by the use of dummy variables or individual regressions. Moreover, in all the pooled regressions the coefficients of current money supply and money supply lagged one year are highly significant and indicate that an increase in the rate of growth of the money supply causes a proportionate increase in the rate of inflation within two years and that the greater part of this adjustment takes place within the first year.

Although the coefficients of current income in all the pooled regressions are significant and have the expected negative sign, an increase in the rate of growth of real income results in a significantly less than proportionate decrease in the rate of inflation. In the alternative lag structures, the addition of lagged income variables as a possible proxy for permanent income does not improve the situation; in fact, the coefficients of income lagged one year are significantly positive rather than negative. However, the intercept terms in the pooled regressions, which indicate significant downward trends in velocity of 1.9 to 6.2 percent per year, may be a more adequate proxy for permanent income than are current and lagged measured income. If the intercept terms are assumed to represent a rising trend in permanent income rather than a falling trend in velocity, then the income coefficients and the intercept terms taken together indicate an almost proportionate relationship between real income and inflation. On the other hand, it may actually be that there is a downward trend in the expected costs of holding real balances which causes the downward trend in velocity, and that the demand for real balances is in fact inelastic with respect to real income.[23]

Studies of the causes of inflation in Latin America are invariably related to the monetarist-structuralist controversy.[24] But, before discussing the implications of the present study for this controversy, the conclusions of Diz, Diaz-Alejandro, and Harberger should be examined. In his

---

countries have different variances in rates of inflation, and these differences tend to be reflected in differences among countries in standard errors of the estimate, as shown in Table 3. Differences in residual variance among countries in the pooled regressions will lead to inefficient estimates of the coefficients in Tables 2 and 4, although the coefficients will be unbiased. To eliminate this possible heteroscedasticity, weighted versions of all the pooled regressions have been run, using as weights for each country's observations the reciprocal of the standard error of the estimate from the corresponding individual regression for that country. The results of the weighted pooled regressions are similar to the results already presented; in particular, the findings on the significances of the coefficients are not changed. However, there are tendencies for the magnitudes of the money supply coefficients to fall and the income coefficients to rise, and this might be expected from the results in Table 3, since low inflation countries now tend to be given more weight.

[23] This would not be inconsistent with the previously reported findings of Diz of income coefficients ranging from −.445 to −.845, and of Diaz-Alejandro of intercepts ranging from −10.5 to −20.2. Moreover, in his recent survey of the demand for money, David Laidler, pp. 90, 106–107, concludes that, although the evidence for economies of scale in money holding is quite tentative, the income elasticity of the demand for money may be as low as 0.8 in the United States for the period after World War II.

[24] For a sample of views on this controversy see Werner Baer and Isaac Kerstenetzky, Harry Johnson, Gerald Meier, and Charles Nisbet. Johnson in particular has emphasized that the controversy includes not only monetary versus structural causes of inflation in Latin America, but also opposing views of stabilization policy, that is, whether the reduction in real income which may be caused by traditional stabilization policies is worse than the costs of inflation itself.

study of Argentine inflation, Diz (especially pp. 121–22) reaches strongly monetarist conclusions, based on his findings that changes in the money supply have a substantial impact on the rate of inflation, while changes in wages or exchange rates do not. Despite roughly similar findings, Diaz-Alejandro (especially pp. 121–24) reaches much more structuralist conclusions. His main argument is that the money supply, which he considers simply a proxy for bank credit, plays only a permissive role in Argentine inflation and is not the underlying cause.[25] Harberger (especially pp. 241–48) reaches a more moderate conclusion, despite his essentially monetarist findings that changes in the money supply dominate changes in wage rates in explaining Chilean inflation, because he also finds that changes in wage rates and money supply are highly correlated. He suggests two possible interpretations: either wages simply transmit inflation, or money is not an exogenous variable as the monetary authority is forced to finance changes in wages.

The studies of Harberger, Diaz-Alejandro, and Diz show that very different conclusions about the monetarist-structuralist controversy can be reached from essentially similar findings. The most important result of the present study for this controversy is that a purely monetarist model, with no structuralist variables, reveals little heterogeneity among Latin American countries, in spite of their extreme diversity. The substantial differences in rates of inflation among these countries cannot under the present model be attributed to structural differences, but must rather be attributed primarily to differences in the behavior of the money supply. The importance of the money supply suggests that further research is needed on its determination, particularly since the studies of Harberger, Diaz-Alejandro, and Diz indicate that the money supply may not be exogenous in every Latin American country.[26] In addition, serial correlation of residuals, slight heterogeneity among countries, and the failure of a purely monetarist model to explain all variations in the rate of inflation suggest that in future research other variables, particularly structuralist variables, might be added to the present model. Even if the causes of inflation in Latin America are adequately explained by the present monetarist model, purely monetarist conclusions may not be implied for stabilization policy, since rates of inflation require up to two years to adjust to changes in the money supply, and this may be a longer period of austerity than most Latin American regimes can tolerate politically.

---

[25] Diaz-Alejandro finds that changes in wage rates are highly correlated with subsequent changes in the money supply (a possibility also mentioned by Diz in a footnote), while changes in the money supply do not have a significant influence on subsequent changes in wage rates. He contends that the high rates of inflation combined with falling real national product in several years since 1949 reflect the existence of cost push inflation, so that the monetary authority is faced with increasing unemployment if it does not permit the money supply to grow in response to increases in wage rates.

[26] Small open countries with fixed exchange rates are, for somewhat different reasons, especially likely not to have exogenous money supplies.

REFERENCES

W. Baer and I. Kerstenetzky, *Inflation and Growth in Latin America*, Homewood 1964.

G. C. Chow, "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, July 1960, *28*, 591–605.

T. E. Davis, "Eight Decades of Inflation in Chile, 1879–1959: A Political Interpretation," *J. Polit. Econ.*, Aug. 1963, *71*, 389–97.

C. F. Diaz-Alejandro, *Exchange-Rate Devaluation in a Semi-Industrial Country: The Experience of Argentina 1955–1961*, Cambridge 1965, 115–25.

A. C. Diz, "Money and Prices in Argentina, 1935–1962," in D. Meiselman, ed., *Varieties*

The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model

By RICHARD DUSANSKY AND PETER J. KALMAN[*]

The phenomenon of "money illusion" has interested economists for some time. It has been studied on both the empirical level and on the theoretical level. In particular, those concerned with the integration of monetary theory and value theory have explored the relationships that exist among the rationale for money demand, the existence of money illusion and monetary economies. (For example, see Robert Clower, Cliff Lloyd, Jacob Marschak, Per Meinich, and Don Patinkin.) Furthermore, the foundations of money illusion are relevant to those who desire nonmoney illusion properties in aggregate models; these researchers need to know the microeconomic restrictions which justify their macroeconomic assumptions. The microeconomic foundations of money illusion, however, are not well established. The purpose of this paper is to set them forth. Within a general neoclassical micro-monetary model of individual utility maximization, where money and commodity prices also enter the objective function, we provide general conditions for the existence of demand functions which are free of money illusion, and present a further testable restriction on demand behavior for this situation.

I

We will consider the general formulation of consumer utility maximization when real balances enter the utility function. Specifically, we consider the problem of maximizing a utility function of the form

$$u = u(x_1, \dots, x_nP) \tag{1}$$

where $x_1, \dots, x_{n-1}$ represent the week's purchases of commodities $1, \dots, n-1$, (we use the traditional Hicksian weekly equilibrium concept), $x_n = m$ is holdings of real money balances at the end of the week, and P is the aggregate price level, subject to a budget constraint

$$y + L = \sum_{i=1}^{n} p_i x_i \tag{2}$$

where y is income, L is the week's initial endowment of money balances, and p_i is the money price of commodity i, $i = 1, \cdots, n-1$ (and where $p_n = 1$.

We assume, following Lloyd, Meinich, and Patinkin, that[1]

$$P = P(p_1, \dots, p_{n-1}) \tag{3}$$

And, for expository convenience, we make the usual assumption that $u(x_1, \dots, x_n)$ has continuous second partials and $P(p_1, \dots, p_{n-1})$ has continuous first partials.[2]

Combining (1) and (3) we now consider the more general problem of maximizing

$$u = f(x_1, \dots, x_{n-1}, m, p_1, \dots, p_{n-1}) \tag{4}$$

with respect to $x_1, \dots, x_{n-1}$ and m subject to (2), where for notational convenience we will write $x = (x_1, \dots, x_{n-1})$,

---

[*] State University of New York at Stony Brook. We thank Don Patinkin and Anthony Santomero for helpful comments.

[1] Meinich and Patinkin use a special case of this assumption. Specifically,

$$P = \sum_{i=1}^{n-1} a_i p_i, \qquad \sum_{i=1}^{n-1} a_i = 1$$

[2] None of our results require these restrictive assumptions.

$p=(p_1, \ldots, p_{n-1})$.[3] This represents the most general treatment of this problem.

The necessary conditions for a local constrained maximum of (4) subject to (2) are:

$$\frac{\partial f(x, m, p)}{\partial x_i} - \lambda p_i = 0, \quad i = 1, \ldots, n-1 \tag{5}$$

$$\frac{\partial f(x, m, p)}{\partial m} - \lambda = 0 \tag{6}$$

$$y + L - \sum_{i=1}^{u} p_i x_i = 0 \tag{7}$$

where λ is a Lagrange multiplier. The sufficient conditions for a local constrained maximum are (5), (6), (7) together with the bordered Hessians alternating in sign (starting with positive), shown as (8).

Under our assumptions we can solve the necessary conditions uniquely for local demand functions which have continuous local first partial derivatives.

These are:

$$x_i = h^i(p, L, y), \quad i = 1, \ldots, n-1 \tag{9}$$

$$m = h^m(p, L, y) \tag{10}$$

We can now consider the issue of money illusion. An individual is said to exhibit money illusion if his demand for commodities is affected by an equi-proportionate change in nominal income, initial nominal balances and commodity prices. More precisely, an individual exhibits money illusion if the demand functions in (9) are *not* homogeneous of degree zero in (p, L, y). The issue of money illusion in demand behavior relates crucially to the first-order conditions (5), (6), and (7). It is necessary to examine the solutions to the first-order conditions and to determine whether these solutions become altered in response to the equi-proportionate change in (p, L, y).

Examination of equations (5), (6), and (7) above reveal that money illusion *generally does exist*; the demand functions in (9) are in general *not* homogeneous of degree zero in (p, L, y). If we allow an equi-proportionate change in (p, L, y), the x_i will in general change in an undefinable

$$\begin{vmatrix} \dfrac{\partial^2 f(x, m, p)}{\partial x_1^2} & \dfrac{\partial^2 f(x, m, p)}{\partial x_1 \partial x_2} & -p_1 \\ \dfrac{\partial^2 f(x, m, p)}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f(x, m, p)}{\partial x_2^2} & -p_2 \\ -p_1 & -p_2 & 0 \end{vmatrix}, \ldots, \tag{8}$$

$$\begin{vmatrix} \dfrac{\partial^2 f(x, m, p)}{\partial x_1^2} & \cdots & \dfrac{\partial^2 f(x, m, p)}{\partial x_1 \partial x_{n-1}} & \dfrac{\partial^2 f(x, m, p)}{\partial x_1 \partial m} & -p_1 \\ \vdots & & \vdots & \vdots & \vdots \\ \dfrac{\partial^2 f(x, m, p)}{\partial x_{n-1} \partial x_1} & & \dfrac{\partial^2 f(x, m, p)}{\partial x_{n-1}^2} & \dfrac{\partial^2 f(x, m, p)}{\partial x_{n-1} \partial m} & -p_{n-1} \\ \dfrac{\partial^2 f(x, m, p)}{\partial m \partial x_1} & & \dfrac{\partial^2 f(x, m, p)}{\partial m \partial x_{n-1}} & \dfrac{\partial^2 f(x, m, p)}{\partial m^2} & -1 \\ -p_1 & \cdots & -p_{n-1} & -1 & 0 \end{vmatrix}$$

---

[3] Clearly, we assume all prices positive, and other variables nonnegative.

fashion. This is due to the fact that the marginal utility functions, $\partial f(x, m, p)/\partial x_i$, in this model depend on the $n-1$ commodity prices, which are now disturbed. Consequently, the marginal utilities will be distinctly affected by the specified disturbance, yielding (in general) new solutions to system (5)–(7).

This can also be seen by simple example. Assume the utility function

$$u = p_1 p_2 x_1 x_2 + x_2 m + m$$

which is a special case of (4). Maximization of this function subject to the budget constraint $y+L=p_1x_1+p_2x_2+m$ yields the following solutions to the first-order conditions:

$$\bar{x}_1 = \frac{p_2^2 + p_2 - (y+L)p_2 + y + L}{p_1(p_2-1)^2}$$

$$\bar{x}_2 = \frac{1}{p_2 - 1}$$

$$\bar{m} = \frac{(y+L)p_2^2 - p_2(y+L) - 2p_2^2}{(p_2-1)^2}$$

$$\bar{\lambda} = \frac{p_2}{p_2 - 1}$$

Now assume that p, L, and y are *doubled*. The utility function now becomes

$$u = 4p_1p_2x_1x_2 + x_2m + m$$

and the constraint becomes

$$2y + 2L = 2p_1x_1 + 2p_2x_2 + m$$

Constrained maximization now yields the following solutions:

$$\hat{x}_1 = \frac{4p_2^2 + 2p_2 + 2(y+L) - 4p_2(y+L)}{2(2p_2-1)^2 p_1}$$

$$\hat{x}_2 = \frac{1}{2p_2 - 1}$$

$$\hat{m} = \frac{8p_2^2(y+L) - 8p_2^2 - 4p_2y}{(2p_2-1)^2}$$

$$\hat{\lambda} = \frac{2p_2}{2p_2 - 1}$$

Clearly, the solutions under the money illusion disturbance $\hat{x}_1$, $\hat{x}_2$, $\hat{m}$, and $\hat{\lambda}$, *differ* from the original solutions $\bar{x}_1$, $\bar{x}_2$, $\bar{m}$ and $\bar{\lambda}$. Consequently, the demand functions do exhibit money illusion; they are *not* homogeneous of degree zero in (p, L, y).

II

In considering a less general framework, Patinkin (pp. 22–30 and 403–405) implicitly suggests that homogeneous of degree zero demand functions result if two assumptions are made. The first assumption requires that the price function, equation (3), be homogeneous of degree one in (p). The second assumption is that the demand for nominal money balances, equation (10), be homogeneous of degree one in (p, L, y). It is of interest to see if these assumptions are sufficient to yield homogeneous of degree zero demand functions in our model. Since P does not appear in any of these equations, this is tantamount to examining the sufficiency of the second assumption alone. It is easy to see that, while in the budget constraint this assumption is consistent with the absence of money illusion, it is *not* consistent with the absence of money illusion in (5) and (6). The marginal utility functions still depend on the $n-1$ commodity prices and will be affected by the equi-proportionate change in (p, L, y). In general, we will get new solutions to system (5)–(7). The demand functions are in general *not* homogeneous of degree zero in (p, L, y).

It is possible, however, to specify conditions under which money illusion will not exist. To do this, it is convenient to rewrite the first-order conditions (5) and (6) in the usual ratio forms:

$$(11)\quad \frac{\dfrac{\partial f(x, m, p)}{\partial x_i}}{\dfrac{\partial f(x, m, p)}{\partial x_j}} = \frac{p_i}{p_j}, \quad i = 1, \ldots, n-1 \quad i \neq j$$

$$(12)\quad \frac{\dfrac{\partial f(x, m, p)}{\partial m}}{\dfrac{\partial f(x, m, p)}{\partial x_j}} = \frac{1}{p_j}$$

The values of the x_i—the demand functions in (9)—will remain invariant if the following two conditions hold:

(i) the marginal utility functions of the nonmonetary commodities are homogeneous of any arbitrary degree k in the money and prices arguments, i.e.,

$$\alpha^k \frac{\partial f(x, m, p)}{\partial x_i} = \frac{\partial f(x, \alpha m, \alpha p)}{\partial x_i}, \qquad \alpha \neq 0,$$
$$i = 1, \ldots, n-1$$

and

(ii) the marginal utility function of the monetary commodity is homogeneous of degree $k-1$ in the money and prices arguments.

To see that conditions (i) and (ii) are sufficient to provide demand functions which are free of money illusion we must examine the first-order conditions, (11), (12), and (7). First, we turn to (11) and (12). Assume there is an equi-proportionate change in (p, L, y) by, say, α. Under (i) the left-hand side of (11) will remain invariant; the right-hand of (11) does not change because the α's cancel out. Under (i) and (ii) equation (12) remains unchanged (each side gets multiplied by $1/\alpha$). Second, we turn to (7). Note that the last term in the sum $\sum p_i x_i$ is m; this means, since α multiplies all terms in (7), the budget constraint remains unaltered if m gets multiplied by α. This occurs only if the m function is homogeneous of degree *one* in (p, L, y).

We now show that conditions (i) and (ii) imply that m is homogeneous of degree one in (p, L, y).[4] Consider the expression

$$(13)\quad A = \sum_{i=1}^{n-1} \frac{\partial m}{\partial p_i} p_i + \frac{\partial m}{\partial L} L + \frac{\partial m}{\partial y} y$$

By Euler's theorem, $m = h^m(p, L, y)$ is homogeneous of degree one in (p, L, y) *if and only if* $A = m$. We will show that this holds if (i) and (ii) are assumed.

From the first-order conditions we can derive expressions for the effects of changes in prices, income, and initial money holdings on m:

$$(14)\quad \frac{\partial m}{\partial p_i} = \lambda \frac{D_{in}}{D} - \sum_{j=1}^{n-1} \frac{\partial^2 f}{\partial x_j p_i} \frac{D_{jn}}{D} - \frac{\partial^2 f}{\partial m \partial p_i} \frac{D_{nn}}{D} + x_i \frac{D_{n+1n}}{D},$$
$$i = 1, \ldots, n-1$$

$$(15)\quad \frac{\partial m}{\partial y} = \frac{\partial m}{\partial L} = -\frac{D_{n+1n}}{D}$$

where D denotes the last bordered Hessian in (8) and D_{ji} is the cofactor of the element in row j column i of its associated matrix.

By using relationships (14) and (15), expression (13) can be rewritten as

$$(16)\quad A = \sum_{i=1}^{n-1} \left[\lambda \frac{D_{in}}{D} - \sum_{j=1}^{n-1} \frac{\partial^2 f}{\partial x_j \partial p_i} \frac{D_{jn}}{D} - \frac{\partial^2 f}{\partial m \partial p_i} \frac{D_{nn}}{D} + x_i \frac{D_{n+1n}}{D} \right] p_i - (L + y) \frac{D_{n+1n}}{D}$$
$$= \frac{1}{D} \left[\lambda \sum_{i=1}^{n-1} p_i D_{in} - \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \cdot \frac{\partial^2 f}{\partial x_j \partial p_i} D_{jn} p_i - D_{nn} \sum_{i=1}^{n-1} \frac{\partial^2 f}{\partial m \partial p_i} p_i + \right.$$

---

[4] We thank Hajime Hori for suggesting this approach.

$$+ D_{n+1n}\left(\sum_{i=1}^{n-1} p_i x_i - L - y\right)\Bigg]$$

Let us consider the extreme right-hand side of equation (16) term by term.

a) $$\lambda \sum_{i=1}^{n-1} p_i D_{in}$$

By a well-known theorem on determinants, the expansion of the elements of the $(n+1)$th column with respect to the cofactors of the nth column of D vanishes. Thus,

$$-\sum_{i=1}^{n-1} p_i D_{in} - D_{nn} = 0$$

Or,

$$\lambda \sum_{i=1}^{n-1} p_i D_{in} = -\lambda D_{nn} \tag{17}$$

b) $$\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} \frac{\partial^2 f}{\partial x_j \partial p_i} D_{jn} p_i$$
$$= \sum_{j=1}^{n-1} D_{jn}\left[\sum_{i=1}^{n-1} \frac{\partial^2 f}{\partial x_j \partial p_i} p_i\right]$$

For each given j, $\partial f/\partial x_j$ is assumed to be homogeneous of degree k in m and p. Thus by Euler's theorem,

$$\sum_{i=1}^{n-1} \frac{\partial^2 f}{\partial x_j \partial p_i} p_i + \frac{\partial^2 f}{\partial x_j \partial m} m = k \frac{\partial f}{\partial x_j}$$

Therefore,

$$\begin{aligned} \sum_{j=1}^{n-1} D_{jn}\left[\sum_{i=1}^{n-1} \frac{\partial^2 f}{\partial x_j \partial p_i} p_i\right] &= \sum_{j=1}^{n-1} D_{jn}\left[k\frac{\partial f}{\partial x_j} - \frac{\partial^2 f}{\partial x_j \partial m} m\right] \\ &= k\sum_{j=1}^{n-1} D_{jn}\frac{\partial f}{\partial x_j} - m\sum_{j=1}^{n-1} D_{jn}\frac{\partial^2 f}{\partial x_j \partial m} \end{aligned} \tag{18}$$

From the first-order conditions we know that

$$\frac{\partial f}{\partial x_j} = \lambda p_j$$

Therefore, using (17), we obtain

$$k\sum_{j=1}^{n-1} D_{jn}\frac{\partial f}{\partial x_j} = \lambda k \sum_{j=1}^{n-1} D_{jn} p_j = -\lambda k D_{nn} \tag{19}$$

Also, by expanding D with respect to the nth column, we obtain

$$D = \sum_{j=1}^{n-1} \frac{\partial^2 f}{\partial x_j \partial m} D_{jn} + \frac{\partial^2 f}{\partial m^2} D_{nn} - D_{n+1n}$$

Namely,

$$\sum_{j=1}^{n-1} \frac{\partial^2 f}{\partial x_j \partial m} D_{jn} = D - \frac{\partial^2 f}{\partial m^2} D_{nn} + D_{n+1n} \tag{20}$$

Substituting (19) and (20) in the extreme right-hand side of equation (18), we obtain

$$\begin{aligned} \sum_{j=1}^{n-1} D_{jn}\left[\sum_{i=1}^{n-1} \frac{\partial^2 f}{\partial x_j \partial p_i} p_i\right] \\ = -\lambda k D_{nn} - mD + m\frac{\partial^2 f}{\partial m^2} D_{nn} - mD_{n+1n} \end{aligned} \tag{21}$$

c) $$D_{nn}\sum_{i=1}^{n-1} \frac{\partial^2 f}{\partial m \partial p_i} p_i$$

$\partial f/\partial m$ is assumed to be homogeneous of degree $k-1$. Thus again by Euler's theorem,

$$\sum_{i=1}^{n-1} \frac{\partial^2 f}{\partial m \partial p_i} p_i + \frac{\partial^2 f}{\partial m^2} m = (k-1)\frac{\partial f}{\partial m}$$

As $\partial f/\partial m = \lambda$ by the first-order condition, we obtain

$$\begin{aligned} D_{nn}\sum_{i=1}^{n-1} \frac{\partial^2 f}{\partial m \partial p_i} p_i \\ = D_{nn}\left[\lambda(k-1) - \frac{\partial^2 f}{\partial m^2} m\right] \end{aligned} \tag{22}$$

Let us substitute equations (14), (17), (21), and (22) in the extreme right-hand side of equation (16). Then we obtain

$$\begin{aligned} A = \frac{1}{D}\Bigg[-\lambda D_{nn} + \lambda k D_{nn} + mD - m\frac{\partial^2 f}{\partial m^2} D_m \\ + mD_{n+1n} - \lambda(k-1) D_{nn} + \end{aligned}$$

$$+ m \frac{\partial^2 f}{\partial m^2} D_{nn} - m D_{n+1n} \Big] = m$$

Thus conditions (i) and (ii) imply that m is homogeneous of degree one in (p, L, y). Conditions (i) and (ii) are indeed sufficient to provide demand functions in (9) which are homogeneous of degree zero in (p, L, y), so that (i) and (ii) characterize the utility functions which yield demand functions that are free of money illusion.[5] Although (i) and (ii) restrict the type of utility functions which can be considered, some interesting functions are still admissible; it can be shown, for example, that Cobb-Douglas functions satisfy (i) and (ii).

III

In this section we develop a further testable result for the case in which the demand functions are free of money illusion. First, however, we discuss the major testable results to which we intend to add. It is important to note that the model employed in this paper is a generalized version of the usual neoclassical model; in addition to commodities, the utility function contains money and prices, thus providing the most general framework, consistent with real balances and commodities in the consumer's objective function, within which to explore the foundations of money illusion. The Slutsky system for this model takes the following form:[6]

$$(23) \quad \frac{\partial x_i}{\partial p_k} + x_k \frac{\partial x_i}{\partial y} = \lambda \frac{D_{ki}}{D} - \sum_{j=1}^{n-1} \frac{D_{ji}}{D} \frac{\partial^2 f}{\partial x_j \partial p_k} - \frac{\partial^2 f}{\partial m \partial p_k} \frac{D_{ni}}{D} + \left(\frac{\partial f}{\partial p_k} \Big/ \lambda \right) \frac{D_{n+1i}}{D} + \frac{\partial x_i}{\partial y} \left(\frac{\partial f}{\partial p_k} \Big/ \lambda \right), \qquad i, k = 1, \ldots, n-1$$

$$(24) \quad \frac{\partial m}{\partial p_k} + x_k \frac{\partial m}{\partial y} = \lambda \frac{D_{kn}}{D} - \sum_{j=1}^{n-1} \frac{D_{jn}}{D} \frac{\partial^2 f}{\partial x_j \partial p_k} - \frac{\partial^2 f}{\partial m \partial p_k} \frac{D_{nn}}{D} + \left(\frac{\partial f}{\partial p_k} \Big/ \lambda \right) \frac{D_{n+1n}}{D} + \frac{\partial m}{\partial y} \left(\frac{\partial f}{\partial p_k} \Big/ \lambda \right), \qquad k = 1, \ldots, n-1$$

As is the custom, the "observables" are grouped on the left-hand side of the equality sign, the "unobservables" on the right-hand side. Comparison of this Slutsky system with the one derived from the usual neoclassical model,

$$\frac{\partial x_i}{\partial p_k} + x_k \frac{\partial x_i}{\partial y} = \lambda \frac{A_{rs}}{A}, \qquad i, k = 1, \ldots, n-1$$

(where A denotes the last bordered Hessian in the second-order conditions corresponding to usual neoclassical model and A_{rs} is the cofactor of the element in row r column s of its associated matrix), reveals the great contrast in the two systems. For example, the right-hand side of (23), the unobservables, consists of much more than the traditional Slutsky-Hicks substitution term. The first term, $\lambda(D_{kn}/D)$, while similar to $\lambda(A_{rs}/A)$ in the neoclassical case, will generally be different. The D_{kn} and D contain second-order partial derivatives of the utility function. Since the utility function in the generalized neoclassical model contains prices and money in addition to the usual consumption commodities, the D_{kn} and D will in general take on different values than A_{rs} and A. In addition, the right-hand side of (23) contains four terms which do not appear at all in the neoclassical case. The second and third terms on the right-hand side represent changes in the marginal utilities of the nonmonetary commodities with respect to price changes, and changes in the marginal utilities of the monetary commodity with respect to price changes. The fourth term, $[(\partial f/\partial p_k)/\lambda]\cdot D_{n+1i}/D$, represents the direct effect on utility due to price changes, and the last term $(\partial x_i/\partial y)[(\partial f/\partial p_k)/\lambda]$, involves multiplicatively the direct effect on utility due to price changes and the income effect. Thus, the unobservables in the general micro-monetary model involve much more than the unobservables (i.e., the usual Slutsky substitution terms) that appear in the standard neoclassical model.

---

[5] Note that, for purposes of deriving homogeneous of degree zero demand functions, it is necessary to treat the marginal utility function of money different than the way in which we treat the nonmonetary marginal utility functions. This is consistent with the fact that the monetary commodity, in this discussion, *is* different, for m is a nominal variable, while the x_i are real variables. But more important the homogeneity property of the demand function for m is dictated by the budget constraint, if the other demand functions are to have the desired homogeneity property.

[6] For the detailed derivation of this system, see our 1972 article, pp. 343–46.

To derive the major predictions of neoclassical theory it is necessary to demonstrate that the right-hand side of (23)–(24) is symmetric and negative semidefinite.[7] As Samuelson has indicated, "The assumption that the form in (96) be symmetrical and negative semidefinite completely exhausts the empirical implications of the utility analysis. All other demand restrictions can be derived from this single assumption. These are bold statements, but they are substantiated by the fact that it is possible to work backwards from the assumption of (96) to an integrable preference field displaying the properties necessary for a maximum" (p. 116). (The matrix in Samuelson's (96) is that of the unobservable terms in the Slutsky equation, traditionally the Slutsky-Hicks substitution terms.) It is possible, with some effort, to show that the right-hand side of the above Slutsky system is symmetric and negative semidefinite.[8] It then follows that *the left-hand side, the observables, also has these joint properties.*

The symmetry property, yields the fundamental equation of symmetry. Focusing on (23),

$$\frac{\partial x_i}{\partial p_k} + x_k\frac{\partial x_i}{\partial y} = \frac{\partial x_k}{\partial p_i} + x_i\frac{\partial x_k}{\partial y}$$

For any two commodities, i and k, it must be true that the price effect on commodity i with respect to the price of commodity k plus the rate of consumption of commodity i times the income effect on commodity i must be equal to the price effect on commodity k with respect to the price of commodity i plus the rate of consumption of commodity i times the income effect on commodity k. This is a testable implication of demand theory; if the equality is observed not to hold, then traditional theory is refuted.

The negative semidefiniteness, along with the assumption of the negativity of the own substitution effect, yields an expression for the traditional demand curve for some commodity i,

$$\frac{\partial x_i}{\partial p_i} < -\, x_i\frac{\partial x_i}{\partial y}$$

which, if x_i is a noninferior good (so that $\partial x_i/\partial y > 0$), says that the demand curve is negatively sloped. Further, defining commodities x_i and x_j to be *complements* if

$$\frac{\partial x_j}{\partial p_i} + x_i\frac{\partial x_j}{\partial y} < 0$$

(and substitutes if this inequality is reversed), one is able to derive a statement about the limitation on the degree of complementarity in the system.[9]

---

[7] It is necessary to show that these properties hold for any arbitrary subset containing $n-1$ of the n variables. One can choose the subset so that it contains the money variable.

[8] See our 1972 article.

[9] For a detailed discussion of the testable implications

To these major testable implications it is possible to add a further restriction on demand behavior for the special case of complete absence of money illusion. Under the assumptions (i) and (ii) above, the demand functions (9) are homogeneous of degree zero in (p, L, y). Using Euler's Theorem on homogeneous functions[10] we may write the following:

$$\sum_{j=1}^{n-1} \frac{\partial x_i}{\partial p_j} p_j + \frac{\partial x_i}{\partial y} y + \frac{\partial x_i}{\partial L} L = 0, \quad i = 1, \ldots, n-1 \tag{25}$$

$$\sum_{j=1}^{n-1} \frac{\partial m}{\partial p_j} p_j + \frac{\partial m}{\partial y} y + \frac{\partial m}{\partial L} L = 0 \tag{26}$$

Multiplying (25) by $1/x_i$, $x_i > 0$ and (26) by $1/m$ we get

$$\sum_{j=1}^{n-1} \frac{\partial x_i}{\partial p_j}\left(\frac{p_j}{x_i}\right) + \frac{\partial x_i}{\partial y}\left(\frac{y}{x_i}\right) + \frac{\partial x_i}{\partial L}\left(\frac{L}{x_i}\right) = 0, \quad i = 1, \ldots, n-1 \tag{27}$$

$$\sum_{j=1}^{n-1} \frac{\partial m}{\partial p_j}\left(\frac{p_j}{m}\right) + \frac{\partial m}{\partial y}\left(\frac{y}{m}\right) + \frac{\partial m}{\partial L}\left(\frac{L}{m}\right) = 0 \tag{28}$$

Since the components of (27) and (28) are elasticities—price elasticities, income elasticities and initial money balance elasticities—we have another direct testable restriction on demand behavior for this model. The prediction that these elasticities must appropriately sum to zero is empirically refutable.

REFERENCES

R. W. Clower, "A Reconsideration of the Microfoundations of Monetary Theory," *Western Econ. J.*, Dec. 1967, *6*, 1–9.

R. Dusansky and P. J. Kalman, "The Real Balance Effect and the Traditional Theory of Consumer Behavior: A Reconciliation," *J. Econ. Theor.*, Dec. 1972, *5*, 336–47.

M. D. Intriligator, *Mathematical Optimization and Economic Theory*, Englewood Cliffs 1971.

C. Lloyd, "The Real Balance Effect and the Slutsky Equation," *J. Polit. Econ.*, June 1964, *71*, 295–99.

C. Lloyd, "Preferences, Separability, and the Patinkin Model," *J. Polit. Econ.*, June 1971, *79*, 642–51.

J. Marschak, "The Rationale of Money Demand and of 'Money Illusion,'" *Metroeconomica*, May-Aug. 1950, *2*, 71–100.

P. Meinich, "Money Illusion and the Real Balance Effect," *Statsökonomisk Tiddskrift*, 1964, *1*, 8–33.

D. Patinkin, *Money, Interest, and Prices*, 2d ed., New York 1965.

P. A. Samuelson, *The Foundations of Economic Analysis*, Cambridge 1947.

---

of neoclassical consumer theory, see M. D. Intriligator, pp. 154–63 and Samuelson, pp. 100–16.

[10] For any function $f(x_1, \ldots, x_n)$ which satisfies

$$\alpha^k f(x_1, \ldots, x_n) = f(\alpha x_1, \ldots, \alpha x_n)$$

we have

$$\sum_{i=1}^{n} \frac{\partial f}{\partial x_i} x_i = kf, \qquad \alpha \neq 0$$

Accounting for the Growth of Output

By SPENCER STAR*

Economists have been able to account for only one-half to two-thirds of the economic growth in the United States.[1] It will be shown in this paper that for the period 1950 to 1960, the large unexplained residual is caused by using aggregated rather than disaggregated data, and by incorrectly using value-added relationships rather than relationships implied by production theory. The use of aggregated data—particularly aggregated labor data—has been responsible for more than one-third of the unexplained residual in *U.S.* manufacturing industries. The use of value-added relations has been the cause of another one-third of the residual growth. By correcting these two types of errors, I have reduced the residual growth over a ten-year period in *U.S.* manufacturing by more than two-thirds, from 47 percent to 13 percent of the annual growth of output.

I. The Way to Measure the Residual

The existence of residual growth is caused by deviations of the real world from a conceptual ideal. This ideal dispenses with nonconstant returns to scale by restricting the phrase "a doubling of all inputs" to mean that a second identical world is taken as the additional inputs. There are no problems with indivisibilities, externalities, or any of the other troublesome real world problems. If, for example, a factory bottles mineral water from a rare spring and one wants to double the inputs, it is done by having a second perfectly identical factory, with its own perfectly identical spring, bottle the water—output must exactly double under these conditions, by definition.

In reality, however, we know such conditions do not exist. Consequently, it is quite possible to double *measured* inputs and to have output less than double. Ideally, we want a measurement ability sophisticated enough to measure every factor that causes a deviation from the ideal of constant returns; then one could see that some of the inputs were in fact not doubled and what was previously labeled decreasing returns to scale would be relabeled diminishing proportional returns.

The same view is taken when considering technological change. Managerial innovations, learning through experience, new techniques, new products, and even creative genius can ideally be treated as measurable inputs to the production function. Presently, this ideal exists only as concept, never as fact. Yet this unattained ideal forms the base upon which we rest the interpretation of residual growth: the residual is a measure of our ignorance, an indicator that tells how far we are from the ideal. Any difference between the growth in measured output and the growth in measured inputs is due to our ignorance, to our inability to measure some of the variables correctly, and to the difficulty in quantifying some of the variables.

Given this framework, it should be apparent that there are many possible causes of a residual, though attention has often focused on *technological progress*, a phrase that has often been used as a synonym for residual growth. This seems to be a poor use of the term. For example, one would certainly want to differentiate between increasing returns to scale and technological progress. Clearly, technological progress is a separate cause of growth. Since I have not tried to measure technological progress, it remains as a possible contributing factor to the final residual. However, my results indicate that its separate contribution is probably quite small especially when compared to the contribution of capital, labor, and the kind of technological change that manifests itself through the changing structure of the labor force.

---

* Visiting assistant professor, University of British Columbia. I would like to thank Robert Hall, Zvi Griliches, and George Borts for their helpful comments.

[1] See Moses Abramovitz, Edward Denison, Dale Jorgenson and Zvi Griliches, and John Kendrick.

I have used the Divisia index to measure the residual. This index can be derived from the assumptions of neoclassical theory, it meets a number of convenient index number criteria,[2] and it is the only index that is invariant to changes in relative prices.[3] By assuming that a continuous, twice differentiable production function exists, that it is homogeneous of degree one, and that all inputs are paid in proportion to their marginal revenue products, one can derive the Divisia index of the residual.

The Derivation

The production function is written as:

$$(1)\quad Y(\mathrm{t}) = F(x_1(\mathrm{t}), x_2(\mathrm{t}), \ldots, x_n(\mathrm{t}); \mathrm{t})$$

Y = output

x_i = the ith input

and t = time

By taking the logarithm of the production function and differentiating with respect to time we get:

$$(2)\quad \dot{Y}/Y = \Big(\sum_1^n F_i\dot{x}_i\Big)/F + F_t/F$$

where a dot over a symbol means the derivative with respect to time; and $F_i = \partial F/\partial x_i$, $F_t = \partial F/\partial \mathrm{t}$. Rearranging and using the relation $x_i/x_i = 1$, we have:

$$(3)\quad F_t/F = \dot{Y}/Y - \sum_1^n (F_i x_i/F)(\dot{x}_i/x_i)$$

Our assumptions allow us to set $F_i x_i/F = \beta_i$, the share of receipts going to the ith input. We have the following partial differential equation defining the growth rate of the index:

$$(4)\quad F_t/F = \dot{Y}/Y - \sum_1^n (\beta_i\dot{x}_i/x_i)$$

This equation is set equal to $\dot{A}/A$, where $A(\mathrm{t})$ is the Divisia index of the residual. It is then rewritten as:

$$(5)\quad \dot{A}/A = \dot{Y}/Y - \sum_1^n (\beta_i\dot{x}_i/x_i)$$

The apparent shift of the function over time, $\dot{A}/A$, is the growth of the Divisia index of the residual. If $\dot{A}/A = .03$, one says the residual is growing at 3 percent. Note that β_i is a variable that depends on time; so are x_i, Y, and F, as well as F_t, $\dot{Y}$, and $\dot{x}_i$.

We can find the definite integral of (5) to get the index of the residual rather than growth rate of the index. Integrating with respect to time from the initial period 0 to the terminal period T gives us:

$$(6)\quad A(\mathrm{T})/A(0) = [Y(\mathrm{T})/Y(0)]\exp\left[-\sum_1^n \int_0^{\mathrm{T}} \beta_i(\dot{x}_i/x_i)\,dt\right]$$

One usually sets $A(0) = 1$ and can thus speak of $A(\mathrm{T})$ being the index of the residual at time T. If $A(2) = 1.06$, the residual has grown 6 percent over two periods or at a rate (roughly) of 3 percent per period.

The conventional method of calculating this index has been to take differential equation (5) and estimate it by using annual growth rates as a discrete approximation to the continuous case. First, annual growth rates of the residual are computed, and then the annual average for a longer period is found and given. Hence, the growth of the index is usually given, not the value of the index.

---

[2] The useful properties of the Divisia index are summarized by Jorgenson and Griliches (1971, pp. 86–88).

[3] See Marcel Richter.

Richter has suggested the following discrete approximation to the index:

$$(7)\quad \Delta A(\mathrm{t})/A(\mathrm{t}) = [\Delta Y(\mathrm{t})/Y(\mathrm{t})] - \sum_{1}^{n} [\beta_i(\mathrm{t})\Delta x_i(\mathrm{t})/x_i(\mathrm{t})]$$

This is the approximation that has been used by Robert Solow, and Jorgenson and Griliches (1967). There has been very little justification of this approximation and no examination of its properties. Instead of Richter's approximation, I have used the following:

$$(8)\quad A(\mathrm{T}) \cong \frac{Y(\mathrm{T})}{Y(0)} \Big/ \prod_{1}^{n} \left(\frac{x_i(\mathrm{T})}{x_i(0)}\right)^{\bar{\beta}_i}$$

where $\bar{\beta}_i = [\beta_i(\mathrm{T}) + \beta_i(0)]/2$

This approximation has no estimation error if the actual mean income share is equal to the estimated share or if each input grows at a constant geometric rate. It does not assume the production function is Cobb-Douglas or *CES* since there is no assumption about the elasticity of substitution. If, however, the elasticity of substitution equals one, then factor shares will be constant and the approximations will be exact. This approximation to the Divisia index is likely to have a small approximation error even when we use data observations that are widely spaced in time. In other words, if one wants to measure the residual between 1950 and 1960, only data for the endpoints—1950 and 1960—are needed for a good approximation of the average annual residual growth. Thus, census data that are available at ten-year intervals can be used with little worry about errors due to the long discrete periods. The derivation of the approximation and its error are presented in a paper by Robert Hall and myself.

II. The Use of Disaggregated Data

Whenever I refer to aggregated data in this paper, I will mean data that have been added together without using any explicit weights. The use of aggregates introduces a possible bias because the input portion of the index becomes:

$$(9)\quad \dot{X}^*/X^* = (\sum_{1}^{n} \beta_i)\,(\sum_{1}^{n} \dot{x}_i/\sum_{1}^{n} x_i)$$

where $\dot{X}^*/X^*$ is the growth rate of the index of aggregated inputs. The correct disaggregated measure is:

$$(10)\quad \dot{X}/X = \sum_{1}^{n} (\beta_i \dot{x}_i/x_i)$$

where $\dot{X}/X$ is the growth rate of the index of disaggregated inputs.

In order to be able to add together different units of items, the items must be homogeneous: each unit must be a perfect substitute for any other unit, i.e., the marginal rate of substitution is constant and the units of measurement are chosen so that the marginal products of every unit are equal. This is clearly not the usual case for inputs. It is therefore to be expected that the use of aggregated data in the Divisia index may cause a bias. The difference between $\dot{X}^*/X^*$ and $\dot{X}/X$ is:

$$(11)\quad \dot{X}^*/X^* - \dot{X}/X = (\sum_{1}^{n} \beta_i)\,(\sum_{1}^{n} \dot{x}_i/\sum_{1}^{n} x_i) - \sum_{1}^{n} (\beta_i \dot{x}_i/x_i)$$

We know that $\beta_i = MP_i x_i/Y$, where MP_i is the marginal product of the ith input. If the inputs are homogeneous, their marginal products are identical. Equation (11) then reduces to zero and there is no bias. Another sufficient criterion for equation (11) to equal zero is that $\dot{x}_i/x_i = \dot{x}_j/x_j$ for all i, j. If neither this condition nor the homogeneity condition is met, then a bias can exist. These results can easily be extended to account for similar biases when using aggregated measures of output.

The theory that has been presented so far is consistent with almost any empirical results. The correction of errors has often been associated with reductions in residual growth, but more careful measurement could conceivably increase the residual. An investigation into growth accounting should include a theory that will predict whether the residual will get smaller or larger understated conditions. The empirical work would then provide a test of the theory. In the next section, I develop the theory further so that the direction and size of the aggregation error can be predicted.

The Error Due to Aggregation

It has already been shown that an error occurs when aggregate data are used. In order to simplify the analysis, let us now consider the aggregation of two different types of labor, L_1 and L_2, and assume that those are the only inputs for the production function. The index using aggregated data is:

$$(12)\quad \dot{A}^*/A^* = (\dot{Y}/Y) - [(\beta_1 + \beta_2) \cdot (\dot{L}_1 + \dot{L}_2)/(L_1 + L_2)]$$

The correct index is:

$$(13)\quad \dot{A}/A = (\dot{Y}/Y) - (\beta_1 \dot{L}_1/L_1) - (\beta_2 L_2/L_2)$$

The difference between the two indexes is a measure of the bias due to aggregation. It is:

$$(14)\quad \dot{A}^*/A^* - \dot{A}/A = -(\beta_1 + \beta_2)[(\dot{L}_1 + \dot{L}_2)/(L_1 + L_2)] + \beta_1 \dot{L}_1/L_1 + \beta_2 \dot{L}_2/L_2$$

Since $\beta_1 + \beta_2 = 1$, equation (14) can be rewritten as:

$$(15)\quad A^*/A^* - \dot{A}/A = W_1(\dot{L}_1/L_1) + W_2(\dot{L}_2/L_2)$$

where

$$W_1 = \beta_1 - [L_1/(L_1 + L_2)]$$

and

$$W_2 = \beta_2 - [L_2/(L_1 + L_2)]$$

And since $W_1 + W_2 = 0$, equation (15) can be further simplified to:

$$(16)\quad \dot{A}^*/A^* - \dot{A}/A = W_1[(\dot{L}_1/L_1) - (\dot{L}_2/L_2)]$$

The bias will be zero if, and only if, $(\dot{L}_1/L_1 - \dot{L}_2/L_2) = 0$ or $W_1 = 0$. The weight W_1 is zero only if $\beta_1 = [L_1/(L_1 + L_2)]$. By definition $\beta_1 \equiv P_1L_1/(P_1L_1 + P_2L_2)$ where P_1 and P_2 are the competitive wages of L_1 and L_2. Hence $W_1 = 0$ only if $P_1 = P_2$, which is the case for homogeneous inputs.

If there is a bias, which direction will it be in? If $P_1 \gtreqless P_2$ then $W_1 \gtreqless 0$. The sign of $(\dot{L}_1/L_1 - \dot{L}_2/L_2)$ depends on the relative growth rates of L_1 and L_2. This produces a rule of disaggregation that will be useful for predicting the effects of disaggregation.

RULE OF DISAGGREGATION: *If, and only if, the prices of the inputs are equal or the growth rates of the inputs are equal, then aggregation does not bias the residual. If there is a bias, disaggregation will decrease the residual when the input with the highest price grows the fastest. It will increase the residual when the input with the lowest price grows the fastest.*

These results are true for any pair of inputs, and if the logic is correct the results are not capable of being falsified. However, when the disaggregation is more complex than merely disaggregating one input to two inputs, the disaggregation rule may be falsified.[4] If it is not disconfirmed by the data, it will provide a useful guide to the bias introduced by using aggregated data.

---

[4] Examples of situations in which the rule is falsified can easily be worked out.

III. Value-Added and Material Inputs

The Divisia index of the residual has been derived from a production function with n inputs and one output. The dependent variable, *output*, is net of any production used within the industry. The total production of the industry including that production used within the industry is *gross production*. In order to be consistent, the material inputs (as compared to capital, labor, and land) need to be divided into two components: materials that come from outside the industry, which I shall call *outside materials*, and materials that are obtained from within the industry, either at the final output stage or at some intermediate stage, which I shall call *inside materials*. Inside materials can then be subtracted from gross production to get output. Usually, however, these data are not available, and one is forced to subtract the total materials (inside and outside) from gross production ending with *value-added* (*VA*). Value-added may also be thought of as output less outside materials. Value-added is used not because it is correct but because it seems to be better than using gross production and because the alternative of using output is simply not usually available.[5] This practice probably is derived in part from the national accounts in which all intermediate goods net out and in part from the lack of a suitable alternative. Fortunately, for the industries and period used in this paper, it was possible to separate inside from outside materials so that output could be calculated.

A comparison between *VA* and output can be made using the framework of the Divisia index. Consider the case with a single output, Y, and two inputs, M and X, where M is units of outside material input and X is a Divisia index of all other inputs. Starting with the production function

$$Y = f(X, M, \mathrm{t}) \tag{17}$$

the Divisia index of the residual is:

$$\dot{A}/A = \dot{Y}/Y - \beta\dot{X}/X - (1-\beta)\dot{M}/M \tag{18}$$

Usually this is incorrectly estimated by substituting $(Y-M)$ for Y and by omitting materials as an input; the entire relative share for M is imputed to the other inputs.

The *VA* index using these substitutions is written:

$$\dot{A}^*/A^* = (\dot{Y} - M)/(Y - M) - \dot{X}/X \tag{19}$$

The difference between the two indexes is a measure of the error due to using *VA* data. Subtracting (19) from (18) results in:

$$\dot{A}/A - \dot{A}^*/A^* = V_1(\dot{Y}/Y - \dot{M}/M) + V_2(\dot{X}/X - \dot{M}/M) \tag{20}$$

where $V_1 = -M/(Y-M)$ and $V_2 = 1-\beta$. If $\dot{Y}/Y = \dot{M}/M$, equation (20) reduces to:

$$\dot{A}^*/A^* - \dot{A}/A = V_2(\dot{Y}/Y - \dot{X}/X) \tag{21}$$

Since $(\dot{Y}-\dot{M})/(Y-M) = \dot{Y}/Y$, the term in parentheses in equation (21) is equal to $\dot{A}^*/A^*$ giving us:[6]

---

[5] For example, Kendrick has written, "Real net output estimates consistent with national income, rather than value-added, would be preferable. Again due to lack of data, we have had to be satisfied with real value-added estimates as a first approximation to the real net measure" (p. 416). Evsey Domar has incorporated materials into a theoretical structure for measuring technological progress, and Lester Lave has attempted to correct some of Solow's estimates of the residual by including materials. M. Ishaq Nadiri has commented that omitting raw materials is another source of bias. Both Nadiri and Lave suggest that the rationale for omitting materials is that the ratio of materials to total output remains constant. Kenneth Arrow has also recently worked on the problem.

[6] It is easily shown that $\dot{Y}/Y = (\dot{Y}-\dot{M})/(Y-M)$ if $\dot{Y}/Y = \dot{M}/M$.

$$\dot{Y}/Y = \left(Y\frac{\dot{Y}}{Y} - M\frac{\dot{M}}{M}\right) \Big/ (Y - M) \tag{a}$$

$$\dot{Y}/Y = (\dot{Y}/Y)(Y - M)/(Y - M) \tag{b}$$

$$(22)\quad (\dot{A}^*/A^* - \dot{A}/A)/(\dot{A}^*/A^*) = V_2 = 1 - \beta$$

or

$$(23)\quad \dot{A}/A = \beta\dot{A}^*/A^*$$

The error due to using VA is given in equation (20). If, as often suggested, the material requirements per unit of output remain relatively stable, then equation (23) provides a good approximation of the error: the growth of the true residual will be some fraction of the VA residual's growth. Since for manufacturing industries the value of β is near one-half, the approximation in (23) suggests the VA residual has been substantially overestimating our ignorance of the growth process.

The existence of a large bias from using VA leads to an important and interesting question. As we have seen, the use of output and materials will eliminate a bias and reduce the residual for a sector, an industry, or a firm. But the inclusion of material inputs has no analogue at the national level: intermediate material inputs are omitted and output is the same as VA. How does one reconcile the effect of introducing output and materials at the firm, industry, or sector level with the lack of such opportunity at the national level?

The simplest way to answer that question is to consider an economy made up of two firms. Firm one produces its output Y_1 from capital, labor and materials, denoted as K_1, L_1, M_2. The materials are output from firm two (all used by firm one) and are thus labeled with the second firm's identifying number. Firm two uses only capital and labor as inputs, denoted as K_2 and L_2. The Divisia indexes of residual growth for the firms are:

$$(24)\quad A_1/A_1 = \dot{Y}_1/Y_1 - \alpha_1\dot{K}_1/K_1 - \beta_1\dot{L}_1/L_1 - \gamma_1\dot{M}_2/M_2$$

$$(25)\quad \dot{A}_2/A_2 = \dot{M}_2/M_2 - \alpha_2\dot{K}_2/K_2 - \beta_2\dot{L}_2/L_2$$

where the Greek symbols are the share of receipts paid to each input. The natural way to combine the two residuals is to use the Divisia index with weights being the relative share of receipts going to each firm:

$$(26)\quad \dot{A}/A = \delta_1\dot{A}_1/A_1 + \delta_2\dot{A}_2/A_2$$

where

$$\delta_1 = Y_1/(Y_1 + M_2)$$

and

$$\delta_2 = M_2/(Y_1 + M_2).$$

This is, however, not the residual obtained when the firms are integrated and the residual is measured at the "national" level. That residual is found by substituting into equation (24) the expression for $\dot{M}_2/M_2$ from (25). The result is:

$$(27)\quad \dot{A}'/A' = \dot{A}_1/A_1 + \gamma_1\dot{A}_2/A_2$$

where $\dot{A}'/A'$ is the growth of the residual measured at the national level. Note that the national residual is a weighted *sum,* not a weighted *average:* the weights[7] add to a number greater than one. As a result of the residuals being additive, at higher levels of integration the residual will be larger. This consequently reconciles the great effect of introducing output and materials at the industry level with the inapplicability of such procedures at the national level.

One loose end still remains. The measure of success in accounting for growth has been not the absolute growth of the residual but rather its growth relative to output. If output growth is defined as the weighted sum $\dot{Y}'/Y' = \dot{Y}_1/Y_1 + \gamma_1\dot{M}_2/M_2$, then $(\dot{A}'/A')/(\dot{Y}'/Y')$ is equal to the measure of our ignorance one gets from using the natural δ weights (from equation (26)) throughout. *The implication of this is that other studies, which use* $(\dot{A}'/A')/(\dot{Y}_1/Y_1)$, *have underestimated our ability to account for growth simply because the wrong denomi-*

---

[7] The weights are the ratio of the value of the firm's output to the value of the nation's output.

nator was used to calculate the proportion explained. It appears that this result has been obscured for so long because growth accounting started from the national level with value-added and worked down to lower levels of integration, instead of starting with the firm's production function and working up to the aggregate level.

IV. The Data

As shown in Section II, aggregation (by adding without explicit weights) of heterogeneous inputs causes a bias in the residual. Consistent data on the labor force at the higher degree of disaggregation desired were available only from the *Census of Population*, 1950 and 1960. The availability of a discrete approximation to the Divisia index (see equation (8)) that needs data at two widely spaced points in time to estimate accurately the average residual growth, meant that data for only 1950 and 1960 could be used with little fear of a large bias being introduced.

Data for capital, labor, and materials were consistently available only for 17 manufacturing industries at the two-digit SIC level.

The capital stocks for 1950 and 1960 were derived by using the estimates of gross capital expenditures made by Michael Gort and Raford Boddy and then adjusting them so that the totals for the industries were consistent with the *OBE* 1966 capital stock study. The price deflator for equipment was developed by Robert Gordon and for structures is the standard Office of Business Economics series.[8] Depreciation rates were based on the lives of stocks as described in *Bulletin F.*[9]

The labor data were collected from the *Census of Population*, 1950 and 1960. The data provide the average weekly hours worked by those working 50 to 52 weeks, cross-classified by industry, sex, occupation, age, schooling, and race.[10]

Data on material inputs came from the 1958 and 1963 *U.S.* input-output tables and from the Office of Business Economics.[11]

Quality Adjustment

The great advantage of using disaggregated data is that *quality changes are transformed into quantity changes.* If a white, male accountant, age 30 (in 1950) with three years of college, who was working full time in the textile industry, got his fourth year of college work done during the 1950's, the labor input for 1960 would show a decline in his former (1950) group and an increase in the corresponding group with four years of college. He also would have been shifted from the 25–34 year old group to the 35–44 year old group. If aggregated data were used, this increase in human capital would have appeared as residual growth. Disaggregation has allowed the identification of the source of the productivity increase.

---

[8] See Gordon, pp. 338–39, col. 1; and *OBE* (1966), Table 8.8, line 2.

[9] See U.S. Bureau of Internal Revenue for *Bulletin F*. Land was not included as part of the capital stock because it was assumed that the amount of land used in manufacturing did not change appreciably during the ten-year period covered. Jorgenson, Griliches, and Denison have all made the same assumption. Inventories contributed to the production of output only through the flow of inventories into the productive process. The net change in inventories of final goods was added to output. The net change in inventories of work in progress and raw materials was usually omitted from the inputs since it was a material input rather than capital or labor. When materials were used as an input, the net changes in inventories were included.

[10] The classes are: (1) industry: SIC 20–39 excluding 24, 29, 28, and 39. SIC 371 was separated from SIC 37; (2) sex: both types; (3) occupation: professional, managerial, clerical, sales, craftsmen, operative, service, laborer; (4) age: 14–24, 25–34, 35–44, 45–54, 55–64, 65+; (5) schooling: (in years) 0–7, 8, 9–11, 12, 13–15, 16+; (6) race: white and nonwhite. This produces 1152 types of labor in each of 17 industries.

[11] See Jack Gottsegen and Richard Ziermer and the *OBE* (1969). I obtained total materials used in the industry from Gottsegen's data; the input-output tables provided the proportion of total materials used that came from within the industry for 1958 and 1963. Then the annual percentage change in the proportions was extrapolated back to 1950 and interpolated for 1960; this gave the proportions for those years. The sensitivity of the results to using only the 1958 and 1963 observations was tested by repeating calculations of the residuals by applying the 1958 proportions to 1950 and 1960. 1960. There were almost no changes in the residuals.

V. Predictions and Empirical Results

The theories developed in Section II are useful not only for the insights that one gains but also because they naturally lead to predictions about the effects of aggregation and the use of *VA*. It is simple to estimate the direction of aggregation bias by looking at manufacturing data to determine if the highest priced inputs grow the fastest. In this section, the predicted effects are compared with the empirical results. In every case the predictions are consistent with the empirical results. For the 17 manufacturing industries studied, *VA* grew at an average annual rate of 3.2 percent.[12] Output grew at a 4.5 percent average annual rate.[13] The ratio of *VA* to aggregated labor grew 2.1 percent annually, or 66 percent as fast as *VA*. When both aggregated capital and labor were used as inputs the average annual residual growth was 1.5 percent—about 47 percent of the average annual growth of *VA*.

Predictions

Starting with the average annual residual growth of 1.5 percent for the 17 industries, the effects of disaggregation can be predicted by using the disaggregation rule.

The capital stock will be disaggregated into two groups, structures and equipment. Structures generally have a significantly longer life span than equipment; hence, $10 of capital structures would rent out at a lower price than would $10 of equipment, i.e., the rental price of structures is less than the rental price of equipment. The *OBE* capital stock study indicates that over the decade the net stock of structures increased 21 percent and the net stock of equipment increased 44 percent.[14] Since the input with the highest price grew fastest, disaggregation should decrease the residual.

The first step in the labor force disaggregation is to classify employees by sex. The number of employed males and females in manufacturing both increased 21 percent over the decade.[15] The predicted result is that disaggregation will have no change on the residual.

Next consider disaggregation by occupation. Table 1 provides information on the growth of the different occupations and gives the mean income. The occupations with the highest incomes grew the fastest with one exception, managers. Perhaps a relative scarcity of managerial talent has caused their wages to be the highest and their growth rate to be only third. Other than that exception, there is a perfect rank correlation between the fastest growing occupations and the highest wages. Since the inputs with the highest prices grew fastest, the predicted result is that disaggregation will reduce the residual.

Disaggregation by age is next. Six age groups are used, starting with 14–24 years and increasing in ten-year intervals. It is well known that workers in the middle years from 35 to 64 are the most productive and highly paid, while those from 14 to 34 and over 65 earn relatively less. A comparison between the growth rates of those in the two large age groups shows

---

[12] A weighted geometric average was used to get the average for all 17 industries. The weights were the ratio of an industry's value-added to the sum of value-added for all 17 industries.

[13] The 4.5 percent is a geometric index of the industry growth rates of output. The weights are the ratio of the industry's output to the sector's output.

[14] Structures data came from Robert Wasson et al., p. 31, Table 5, series II. Equipment data came from the same study, p. 34, Table 6.

[15] 1960 data: U.S. Bureau of the Census (1963a), Table 213, pp. 1–569 to 1–570. 1950 data: U.S. Bureau of the Census (1953), Table 133, pp. 1-288 to 1–289.

TABLE 1—EMPLOYMENT GROWTH, BY OCCUPATION FOR MALES EMPLOYED IN MANUFACTURING, 1950 TO 1960

Occupation	Mean Wage and Salary Income 1960	Ratio of Number of Males in Each Occupation, 1960/1950
Professionals	8,590	1.95
Sales	6,739	1.51
Managers	9,870	1.26
Craftsmen	5,825	1.21
Clerical	5,336	1.19
Operatives	4,946	1.13
Services	4,135	0.98
Laborers	4,101	0.81

Sources: 1960 Data: U.S. Bureau of the Census (1936a), Table 208, pp. 1-557 to 1-562. 1950 Data: U.S. Bureau of the Census (1953), Table 134, pp. 1-289 to 1-290.

that the high-pay group (ages 35–64) grew 36 percent while the low-pay group grew only 2 percent.[16] Disaggregation should therefore substantially reduce the residual.

The next candidate for disaggregation is education. It should be noted that I measure only formal schooling and loosely refer to it as education. Numerous studies have shown that there is a positive private monetary return for schooling. Other things being equal, the more schooling one has, the higher is one's income. Hence, if those groups with the most years of schooling grew the fastest, we would, according to our rule, expect disaggregation to decrease the residual growth. This in fact occurred as can be seen by referring to Table 2.[17] Consequently, disaggregation should decrease the residual.

The last disaggregation is by race. For all of manufacturing employment from 1950 to 1960, the number of nonwhites increased 20.6 percent while the number of whites increased 20.2 percent.[18] The residual should be very slightly increased by disaggregation since the lowest paid group, nonwhites, grew the fastest.

TABLE 2—CHANGES IN SCHOOLING BY EXPERIENCED LABOR FORCE, MALES, AGE 25-64, 1950 TO 1960

Years of School Completed	Ratio 1960/1950
0–7	0.79
8	0.88
9–11	1.25
12	1.30
13–15	1.48
16+	1.55

Sources: 1960 Data: U.S. Bureau of the Census (1963a)' Table 9, pp. 116–129. 1950 Data: U.S. Bureau of the Census (1956), Table 10, pp. 1B–107 to 1B–108.

Having considered the bias introduced by aggregation, we can turn to the bias that occurs from using *VA* and from omitting materials as inputs. As noted in Section III, if the ratio of materials to output is constant, the growth of the true residual will be a fraction of the *VA* residual's growth, the fraction being the share of receipts going to all nonmaterial inputs. For the manufacturing industries studied, that share is somewhat over one-half. Consequently, elimination of this source of bias should reduce the residual growth substantially, and if materials have remained a relatively constant proportion of output, the mean residual growth should be reduced roughly by 50 percent.

To summarize the predictions, they indicate that disaggregation will substantially reduce the residual in all cases except for disaggregation by sex and by race. Disaggregation by sex should have no effect while disaggregation by race should slightly increase the residual. Furthermore,

---

[16] 1960 data: U.S. Bureau of the Census (1963a), Table 212, pp. 1-567 to 1-568. 1950 data: U.S. Bureau of the Census (1953), Table 132, pp. 1-286 to 1-287.

[17] The data are for the occupation being studied, but are for males, ages 25–64, in the entire United States. There is little reason to believe that data for manufacturing, for males and females, over the entire age span, would be substantially different.

[18] 1960: White, 16,277,720; Nonwhite, 1,235,366. 1950: White, 13,545,647; Nonwhite, 1,025,104. 1960 data: U.S. Bureau of the Census (1963a), Table 91, pp. 1-221 to 1-222. 1950 data: U.S. Bureau of the Census (1953), Table 133.

TABLE 3—RESIDUALS FOR 17 MANUFACTURING INDUSTRIES, 1950–1960

Inputs	Average Annual Residual Growth In Percent	Growth of Residual As Percent of Annual Growth of Value-Added
K, L	1.51	47
$K(EQ, ST), L$	1.41	44
$K(EQ, ST), L(S)$	1.42	44
$K(EQ, ST), L(S, O)$	1.21	38
$K(EQ, ST), L(S, O, A)$	1.04	32
$K(EQ, ST), L(S, O, A, E)$	0.82	26
$K(EQ, ST), L(S, O, A, E, R)$	0.87	27
$K(EQ, ST), L(S, O, A, E, R), M$	0.59[a]	13[b]

Note: Abbreviations in parentheses above refer to the disaggregated factors. L=Labor Hours; Disaggregated by S=Sex, O=Occupation, A=Age, E=Education, R=Race; K=Net Capital Stock; Disaggregated by EQ=Equipment, ST=Structures; M=Materials from outside the industry.

[a] This is the residual for the manufacturing sector using the ratios of the industry's output to the sector's output as weights.

[b] Expressed as a percent of the growth of output, not value-added.

the adjustment of the residual measurement to include materials and to use output rather than value-added will reduce the correct residual to some fraction of the value-added residual's growth.

Empirical Results

The results are presented in Table 3. Using only aggregated capital and aggregated labor as inputs and VA as output for each of the 17 industries, the average annual residual growth is 1.51 percent, which is 47 percent of the growth rate of VA.

Disaggregation of the capital stock by structures and equipment reduced the residual growth from 47 percent to 44 percent of VA growth. The relatively small effect is due mainly to the small weight capital receives relative to the labor input. When the effect of the different weights is considered, then the reduction of the residual from separating structures and equipment is seen to be substantial.

A warning is necessary before discussing the results of disaggregating the labor input. The results will be presented in steps; the reduction in residual growth is presented after each level of disaggregation has taken place. The numerical results achieved depend on the particular order in which the data are disaggregated. A simplified example may make this clear. Suppose that there are two sexes, men and women, and that men are educated while women are not. If one disaggregated by sex first and then by education, all the improvement in the residual appears to be caused by the differences between the sexes. Education "contributes" nothing to the reduction of the residual. If the order of disaggregation is reversed, the opposite result appears. In other words, to the extent that the "causes of growth" move together, the first one chosen will be labeled as being the "cause" or the "explanation" of growth. Keeping in mind this caveat, let us go on to consider the results from disaggregating labor.

As expected, the disaggregation by sex had almost no effect on the residual, changing the average annual rate only from 1.41 to 1.42 and leaving the residual growth at 44 percent of the growth of VA. When the labor force was disaggregated by the eight occupational groups, the residual growth was reduced to an annual rate of

1.21 percent, about 38 percent of the *VA* growth rate.

Another large reduction in the residual's growth rate took place when the labor force was divided into six age groups, the residual now growing only 32 percent as fast as *VA*. The structural changes in the age of the labor force have been a significant source of the previously unaccounted for residual growth.

When the industry, sex, occupation, age cells were further disaggregated by education, the residual once again had a large decline. It dropped from an annual percentage rate of 1.04 to 0.82 decreasing the residual growth from 32 percent to 26 percent of the *VA* growth. This is even more impressive when one realizes that all the correlated effects from sex, industry, occupation, and age have already been removed. If education had been one of the first steps in disaggregation, the size of the change would undoubtedly have been much larger. Race was the last disaggregation to be performed. As predicted, the annual residual growth increased—from 0.82 to 0.87 percent or from 26 to 27 percent of the growth of *VA*. The extremely small differences in the relative growth rates appear to provide a good test of the sensitivity of the disaggregation rule.

We thus end the disaggregation having reduced the residual from 47 percent (with aggregated capital and labor) to 27 percent of output growth. The 20 percentage point reduction of the residual growth means that 43 percent of the unexplained *VA* growth has now been accounted for. The disaggregation rule has been consistently verified by the empirical results, even in the most complex and difficult case.

After all the disaggregation had been done, I replaced *VA* with output, included materials as input, adjusted the weights, and recalculated the residual for each two-digit industry. The annual residual growth was reduced from 27 percent of the growth of *VA* to 13 percent of the growth of output, the direction being that suggested by the theory. The geometric mean of the residual growth for the 17 industries was 0.037, which is about 43 percent of the value-added residual. Since materials are about half the value of output, equation (23), which is based on the assumption of a constant materials output ratio, turns out being a slightly inaccurate predictor. The implication is that the materials output ratios have varied but not dramatically.

VI. Concluding Comments

The major sources of economic growth in manufacturing during the 1950's were increases in capital, increases in the number of employees, and structural changes that affected the occupational, age, and educational makeup of the labor force. Changes in the sexual and racial structure were very slight and thus had no substantial impact on the growth of output. The isolation of these elements as the sources of growth was due to the combination of having extremely disaggregated data and having a practical rule of thumb for estimating the effect of disaggregation. Just as important as pinpointing actual sources of economic growth was the discovery of a spurious cause of unexplained economic growth, the bias introduced by using *VA* and omitting materials. The correction of this source of bias also led to the finding that an inconsistent measure of ignorance is usually used, one that takes the ratio of a weighted sum to a weighted average. This measure was replaced by one that is more natural and consistent.

The techniques used throughout this paper are, of course, applicable to the entire economy. They also could be applied even more intensively than I was able to do here. I expect that finer disaggregation of the capital stock, of output, and of the occupational component of the labor force would lower the residual further; and disintegration down to the level of a four-digit SIC industry would most likely eliminate the residual all together, except for the probably minor effects of inaccurate statistics, omitted inputs, and returns to scale.

As a final word, something should be said about technological progress. Neutral disembodied technological change remains as a possible unaccounted for contribution to the growth of output; however, it can only play a very limited part in the growth of output since the residual is so small. In a broader sense, though, technological change is deeply involved behind the scene. The great increase in the number of relatively highly paid employees has been responsible for a significant proportion of the growth in manufacturing. Shifts in the demand and supply curves in different labor markets lie behind these structural changes, and technological progress plays a very important role in explaining the causes of these shifts. The introduction of new products and new techniques increases the demand for physical and human capital while education and research cause shifts in the supply. Consequently, the reduction of residual growth does not eliminate technological progress as an important causal factor in growth. The changing structure of the labor force, a factor emphasized in this paper, is a result in large part of the forces of technological progress. The residual is a measure of our ignorance about only one aspect of growth. Growth accounting measures *what* changes have taken place. The most interesting question—*why* the changes occurred—still remains to be answered.

REFERENCES

M. Abramovitz, "Resource and Output Trends in the United States since 1870," *Amer. Econ. Rev. Proc.*, May 1956, *46*, 5–23.

J. K. Arrow, "The Measurement of Real Value-Added," tech. rept. no. 60, Econ. Series, Institute for Mathematical Studies in the Social Sciences, Stanford Univ., June 1972.

E. F. Denison, *The Sources of Economic Growth in the United States*, supp. pap. no. 13, Committee for Economic Development, Washington 1962.

———, *Why Growth Rates Differ*, Washington 1967.

E. Domar, "On the Measurement of Technological Change," *Econ. J.*, Dec. 1961, *71*, 709–29.

R. J. Gordon, "Problems in the Measurement of Real Investment in the U.S. Private Economy," unpublished doctoral dissertation, Harvard Univ. 1968.

M. Gort and R. Boddy, "Gross Capital Expenditures," Memo. 1968.

J. J. Gottsegen and R. C. Ziermer, "Comparison of FRB and OBE Measures of Output of Real Manufacturing Output, 1947–64," in J. Kendrick, ed., *The Industrial Composition of Income and Product*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 32, New York 1968.

D. Jorgenson and Z. Griliches, "The Explanation of Productivity Change," *Rev. Econ. Stud.*, July 1967, *34*, 249–83.

——— and ———, "Issues in Growth Accounting: A Reply to Edward F. Denison," dis. pap. no. 215, Harvard Institute of Economic Research, Nov. 1971.

J. W. Kendrick, *Productivity Trends in the United States*, Princeton 1961.

L. Lave, *Technological Change: Its Conception and Measurement*, Englewood Cliffs 1966.

M. I. Nadiri, "Some Approaches to the Theory and Measurement of Total Factor Productivity: A Survey," *J. Econ. Lit.*, Dec. 1970, *8*, 1137–77.

M. K. Richter, "Invariance Axioms and Economic Indexes," *Econometrica*, Oct. 1966, *34*, 739–55.

R. M. Solow, "Technical Change and the Aggregate Production Function," *Rev. Econ. Statist.*, Aug. 1957, *39*, 312–20.

S. Star and R. E. Hall, "An Approximate Divisia Index of Total Factor Productivity," disc. pap. no. 73-19, Univ. British Columbia, Oct. 1973.

Robert C. Wasson, J. C. Musgrave, and C. Harkins, "Alternative Estimates of Fixed Business Capital in the United States, 1925–1968," *Surv. Curr. Bus., 50,* no. 4, Apr. 1970, 18–36.

U.S. Bureau of the Census, *U.S. Census of Population: 1950,* Vol. II, *Characteristics of the Population,* Part 1, U.S. Summary, Washington 1953.

———, *U.S. Census of Population: 1950,* Vol. IV, Special Reports, P-E No. 1B, Part 1, ch. B, *Occupational Characteristics,* Washington 1956.

———, (1963a) *U.S. Census of Population: 1960,* Vol. I, *Characteristics of the Population,* Part I, U.S. Summary, Washington 1963.

———, (1963b) *U.S. Census of Population: 1960,* Subject Reports, *Occupational Characteristics,* Final Report PC(2)-7B, Washington 1963.

U.S. Bureau of Internal Revenue, *Income Tax Depreciation and Obsolescence, Estimated Useful Lives and Depreciation Rates,* rev. ed., Washington 1942.

U.S. Department of Commerce, Office of Business Economics, *(OBE), National Income and Product Accounts: 1929–1965, Surv. Curr. Bus.* supp. 1966.

———, National Economics Division, "Input-Output Structure of the U.S. Economy: 1963," *Surv. Curr. Bus.,* Nov. 1969, *49,* no. 11.

The Stability Properties of Optimal Economic Policies

By STEPHEN J. TURNOVSKY*

Recent writings on optimal stabilization policy have drawn attention to the problem of "instrument instability."[1] This term describes a problem which may be encountered by a policy maker when, in order to optimally stabilize a set of chosen target variables, he is required to continually increase the magnitude of the adjustment of his policy instruments, so as to offset the effects of past policies. This is an important consideration since it indicates a very real difficulty which may arise if the government attempts to "fine tune" the economy. The first detailed description of this question is contained in a recent paper by Holbrook. As he points out, the nature of the problem is intimately bound up with the weighting pattern describing the impact of present and past values of the policy instruments on current values of the target variables.

Much of Holbrook's paper is concerned with using a very simple model to show how quite plausible lag patterns may lead to instrument instability and with using these results to test some available empirical evidence. The purpose of the present paper is to extend his analysis in two directions: first to examine the stability problem within the context of a more general dynamic model and secondly, to allow the coefficients in the system to be subject to random disturbances. As will be seen below, both these extensions, and in particular the latter, yield considerable insight into the likely occurrence of the problem.

To do this, we must make use of some results recently developed in modern deterministic and stochastic control theory.[2] These results, which appear almost entirely in the engineering literature, have been developed primarily for continuous time systems. The analogous results for discrete time systems—which are more relevant for problems of economic stabilization—have surprisingly been given very little attention. The only statement of the relevant theorems for discrete time stochastic systems of which I am aware, is rather inaccessible, being contained in some unpublished lecture notes by W. M. Wonham.[3] Since the discrete time results are substantively different from those obtained for continuous time systems, and since they are in fact probably of more interest for economists, this seems an appropriate occasion to outline these general solutions. Thus a second objective of this paper is largely expository, being to restate in a form convenient for economists, some important results derived elsewhere by control theorists.[4] While our exposition is rather general, it is clear that these results are applicable to a wide variety of economic stabilization problems.

---

* Professor of economics, Australian National University. I wish to thank Murray Wonham for many helpful discussions related to stochastic control theory.

[1] See Edward Gramlich and Robert Holbrook.

[2] Gregory Chow also extends Holbrook's work by reconsidering it from the point of view of optimal control theory. The present paper differs from his in several important respects in terms of such things as the detailed formulation of the problem, the precise issues considered, and the general emphasis of the discussion. One difference worth mentioning explicity is that the present paper considers the case where the coefficients of the system are stochastic. As shown below this turns out to have an extremely important bearing on the likely occurrence of instrument instability.

[3] The case of the discrete time system with additive disturbances is treated by K. J. Astrom and Chow (1970, 1972). Wonham's notes deal with the more general case where the parameters themselves are also allowed to contain stochastic components. Wonham has published the analogous results for continuous time systems, but as mentioned in the text these lead to substantively different results from those obtained using discrete time.

Making use of these theorems we are able to derive the following results. First, for deterministic systems, it is shown that as long as the policy maker attributes positive costs to both target disequilibrium and instrument control, then the optimal policy will ensure both target stability and instrument stability. This result, expressed in somewhat different language, is well known to control theorists. Secondly, assuming that there are no instrument costs (the case considered by Holbrook and others) then instrument instability can arise, although target stability still exists. However, even with zero instrument costs, instrument instability is shown to be impossible for a very important class of dynamic models.

The analogous results are obtained for systems in which the coefficients are subject to stochastic disturbances. It is shown how the presence of these disturbances tends to reduce the phenomenon of instrument instability. To illustrate this, we take the stochastic analogue of the deterministic equation studied by Holbrook, in which with zero control costs instability could easily arise, and show that with just one extreme exception the policy adjustment will now always be stochastically stable in a very strict sense. Thus, since typically the policy maker is operating in an environment in which parameters are stochastic, these results suggest that the phenomenon of instrument instability is in fact much less of a problem than a purely deterministic analysis would lead us to believe.

I. Stability in a Deterministic Model

Consider the following matrix equation describing a set of relationships between the current values of a vector of target variables y_t, its past values distributed over m previous periods, a vector of policy instruments or control variables u_t, and also its values distributed over the previous n periods:

$$(1)\quad y_t = A_1 y_{t-1} + A_2 y_{t-2} + \dots + A_m y_{t-m} + B_0 u_t + B_1 u_{t-1} + \dots + B_n u_{t-n}$$

Assume that there are q_1 target variables and q_2 policy instruments so that the A_i matrices are square of order q_1 and the B_j matrices are $q_1 x q_2$. For the present we need place no restrictions on the relative numbers of targets and instruments. In addition there may be other exogenous variables not subject to government control, but since these are not essential to the analysis we need not consider them here.

Assume further that the policy maker has in mind some long-run desired levels of the target variables, $\bar{y}$, together with a set of long-run desired values of the policy instruments, $\bar{u}$, which he wishes to attain. These two sets of long-run objectives chosen by the policy maker are assumed to be actually attainable by the system. Consequently they are required to satisfy equation (1) so that:

$$(2)\qquad \left(I - \sum_{i=1}^{m} A_i\right)\bar{y} = \sum_{i=0}^{n} B_i \bar{u}$$

Note that if $q_1 = q_2$ and the matrix $\sum_{i=0}^{n} B_i$ is nonsingular, then for given target objectives $\bar{y}$, equation (2) defines a unique long-run policy $\bar{u}$. Otherwise, this equation provides a set of conditions which the desired values of the target and instrument variables, however chosen, must satisfy in order to be consistent with long-run equilibrium.[5]

---

[4] I am grateful to Wonham for access to his material and for permission to use it. There are two differences between the way he develops his results and the way I report them here. First, he uses notation, which although convenient for his purposes, is cumbersome to develop. The results stated in Section II are expressed in terms more familiar to economists. Secondly, I have generalized these results to permit more disturbances to be correlated.

The objective of the policy maker is to control the economy so as to keep the target variables as close to equilibrium as possible. Following previous authors, we shall assume that the costs of being away from equilibrium can be adequately measured by the quadratic cost function,[6]

$$C = \sum_{t=1}^{\infty} (y_t - \bar{y})' M (y_t - \bar{y}) \tag{3}$$

where the prime denotes vector transpose. The cost matrix M is taken to be strictly positive definite, meaning that the policy maker attributes independent costs to each of the target variables.[7]

In addition there may be control costs associated with the instruments. In many cases, such as those associated with government expenditure programs, the nature of these costs is clear enough. But in other cases, such as in monetary policy, the rationale for including control costs is by no means clear. For example, there are no obvious direct costs associated with having the rate of interest (taken as an instrument variable) away from its long-run equilibrium. What costs are incurred are more likely to be felt via the effects on other target variables. However, we shall commence with the case where all instruments have costs associated with them so that total costs become

$$\sum_{t=1}^{\infty} (y_t - \bar{y})' M (y_t - \bar{y}) + \sum_{t=1}^{\infty} (u_t - \bar{u})' N (u_t - \bar{u}) \tag{4}$$

where N is also strictly positive definite. Presently, we shall also deal with the case considered by Holbrook and others where $N=0$. The intermediate case where there are only some instrument costs (so that N is positive semidefinite) is covered by the latter case.

Subtracting (2) from (1) we can describe the dynamics in deviation form as:

$$\begin{aligned} y_t - \bar{y} = {} & A_1(y_{t-1} - \bar{y}) + \dots \\ & + A_m(y_{t-m} - \bar{y}) \\ & + B_0(u_t - \bar{u}) + \dots \\ & + B_n(u_{t-n} - \bar{u}) \end{aligned} \tag{5}$$

Letting y_t now stand for the *deviation* $y_t - \bar{y}$ and similarly letting u_t denote the deviation $u_t - \bar{u}$, the optimal stabilization problem can now be expressed more conveniently as being to minimize:

$$\sum_{t=1}^{\infty} [y_t' M y_t + u_t' N u_t] \tag{6a}$$

subject to

$$\begin{aligned} y_t = {} & A_1 y_{t-1} + \dots + A_m y_{t-m} \\ & + B_0 u_t + \dots + B_n u_{t-n} \end{aligned} \tag{6b}$$

where henceforth y_t and u_t refer to *deviations* from their respective long-run values.

In order to apply the relevant results of control theory to this problem it is convenient to transform equation (6b) into an equivalent first-order system. Letting:

---

[5] Following the usual stabilization literature we shall assume that the long-run values of $\bar{y}$ and $\bar{u}$ are given. However, the question of how these equilibrium values are determined is also of some importance and should not be ignored.

[6] The limitations of the quadratic cost function are well known. Its main advantage is that it leads to linear optimal decision rules and this is very convenient in view of the linearity of our basic dynamic relationships.

[7] For example if

$$M = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

then

$$C = \sum_{t=1}^{\infty} (y_{1t} - y_{2t})^2$$

where y_{1t}, y_{2t} are the components of y_t. M is thus only positive semidefinite and the costs are defined in terms of the differences between those two components and not in terms of each independently.

$$(7)\qquad x_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-(m-1)} \\ u_t \\ \vdots \\ u_{t-(n-1)} \end{bmatrix}, \qquad v_t = u_t,$$

$$A = \begin{bmatrix} A_1 & A_2 \ldots A_m & B_1 \ldots B_n \\ I & 0 \ldots 0 & 0 \ldots 0 \\ \vdots & \vdots & \vdots \quad \vdots \\ 0 & 0 \ldots I & 0 \ldots 0 \\ 0 & 0 \ldots 0 & 0 \ldots 0 \\ 0 & 0 \ldots 0 & I \ldots 0 \\ \vdots & \vdots & \vdots \quad \vdots \\ 0 & 0 \ldots 0 & 0 \ldots I \end{bmatrix}$$

$$B = \begin{bmatrix} B_0 \\ \vdots \\ 0 \\ I \\ \vdots \\ 0 \end{bmatrix}$$

the system can be written as

$$(8)\qquad x_t = Ax_{t-1} + Bv_t$$

where x_t is the state vector and v_t denotes the control. Note that x_t includes u_t, so that the introduction of the notation v_t is made to avoid ambiguity.

Before we can proceed, we must add one further assumption, namely that the system (8) is fully controllable. By this it is meant that it is always possible to move the system from any one state to any other and is clearly a prerequisite for any optimal control. The condition for full controllability is that the rank of the matrix $[B, AB, \ldots, A^{g-1}B] = g$, where $g = mq_1 + nq_2$, the number of components in x_t (see M. Athans and P. Falb, p. 205).[8]

It remains to rewrite the cost term (6a) in terms of the new notation. To do this

$$\text{let } \tilde{M} = M/m \qquad \tilde{N} = N/n$$

and define

$$(9)\qquad M^* = \begin{bmatrix} \tilde{M} & 0 \ldots\ldots\ldots\ldots 0 \\ 0 & \tilde{M} \ldots\ldots\ldots\ldots 0 \\ \vdots & \vdots \qquad \vdots \qquad\qquad \vdots \\ 0 & . \quad \tilde{M} \ldots\ldots 0 \\ 0 & . \quad . \quad \tilde{N} \ldots 0 \\ \vdots & \vdots \quad \vdots \quad \vdots \qquad \vdots \\ 0 & 0 \quad 0 \quad 0 \ldots \tilde{N} \end{bmatrix} \equiv \begin{bmatrix} M_1^* & 0 \\ 0 & M_2^* \end{bmatrix}$$

which is a block diagonal matrix with m blocks of $\tilde{M}$ and n blocks of $\tilde{N}$. It is clear that M^* is strictly positive definite if, and only if, *both* M and N are strictly positive definite. With some manipulation, the cost term (6a) can be written in the form:

$$(10)\qquad \sum_{t=1}^{\infty} x_t' M^* x_t + \sum_{t=1}^{\infty} v_t' N v_t - \theta$$

where θ includes terms in y_t and u_t occurring before time $t=1$.[9] Since we are dealing with costs incurred for $t \geqq 1$ we need not be concerned with the costs summarized by θ. Thus we can effectively postulate the stabilization problem as being to minimize

$$(11)\qquad \sum_{t=1}^{\infty} x_t' M^* x_t + \sum_{t=1}^{\infty} v_t' N v_t$$

subject to

$$(8)\qquad x_t = Ax_{t-1} + Bv_t$$

---

[8] The conditions in Athans and Falb are given for continuous time systems, but they are identical in the discrete case.

[9] This can easily be seen by multiplying out the cost term and regrouping all the terms in y_t, u_t occurring before $t=1$ into θ.

This is now in standard form and we have the result that if both M^* and N are strictly positive definite, then the optimal policy exists, is unique, and is given by:[10]

$$v_t = Kx_{t-1} \tag{12a}$$

where

$$K = -[N + B'SB]^{-1}B'SA \tag{12b}$$

and S is the unique strictly positive definite solution to the matrix equation:[11]

$$S = M^* + K'NK + (A + BK)'S(A + BK) \tag{12c}$$

The set of equations (12) yields a feedback law whereby the policy maker may compute the optimal policy v_t given the immediate past state of the system. Thus recalling the definition of x_t, the optimal policy $v_t(=u_t)$ is of the form[12]

$$u_t = \sum_{i=1}^{m} \alpha_i y_{t-i} + \sum_{j=1}^{n} \beta_j u_{t-j}$$

where the α_i and β_j can in principle be computed from equations (12b) and (12c).[13]

The most relevant issue in the present context is the stability properties of the system, when the optimal law (12) is applied. Assuming both M and N (and hence M^* and N) are strictly positive definite, we can readily establish the stability of *both* targets and instruments in this case.

Since S is strictly positive definite,

$$V_t \equiv x_t' S x_t$$

is a positive function and thus can serve as a Liapunov function.[14] Accordingly, the vector x_t will tend to zero provided

$$V_t < V_{t-1} \qquad \text{for all t}$$

Substituting from (12a) we see that when the optimal policy is applied, the state variable follows the equation

$$x_t = (A + BK)x_{t-1} \tag{13}$$

Hence using (12c) we find

$$\begin{aligned} V_t - V_{t-1} &= x_{t-1}'[(A + BK)'S(A + BK) - S]x_{t-1} \\ &= -x_{t-1}'[M^* + K'NK]x_{t-1} < 0 \end{aligned}$$

thus implying that $x_t \to 0$. But since x_t includes *both* y_t and u_t (as well as their respective lags), it follows that both these vectors must tend to zero. Hence we can conclude that as long as there are costs ascribed to all targets and instruments, then the optimal policy will ensure the stability of both targets and instruments. Note, however, that this does *not* ensure that all targets can be stabilized at their equilibrium values for *each* period. With control costs present this may be too costly to do. Even without them it will be infeasible if there are more targets than instruments.

By substituting (12a) into (11), performing the summation and utilizing

---

[10] This result is due to Wonham, although similar results have been obtained by Astrom.

[11] Essentially the proof establishes that there exists a positive semidefinite matrix S which is the solution to equation (12c). Having established the existence of such a matrix, the strict positive definiteness of M alone ensures that S is unique and in fact strictly positive definite, even if $N=0$. If $N=0$ and M is only positive semidefinite, then S is also only positive semidefinite and in fact may no longer be unique, as shown by the example in Section III.

[12] A linear feedback policy is still obtained even if the matrices A, B, M^*, N are functions of time and the time horizon is finite. In that case S is the positive definite solution to the equation $S_{t-1}=M_t^*+K_t'N_tK_t+(A_t+B_tK_t)'S_t(A_t+B_tK_t)$ with K_t defined as above and where S_t must satisfy appropriate boundary conditions. The same stability results apply in this case, although explicit analytical solution of the optimal policy becomes much more difficult.

[13] Where there are more than two state variables and two policy variables, explicit analytical solution of equations (12) becomes virtually impossible and it becomes necessary to resort to numerical approximations (see Wonham and Cashman).

[14] A function $V(z)$ is positive if, and only if, $V(z)>0$ for $z\neq 0$, and $V(0)=0$. For a discussion of Liapunov stability see Joseph Lasalle and Solomon Lefschetz.

(12c), it can readily be shown that the total costs incurred when the optimal policy is applied starting from an initial state x_0, are given by the expression:

$$C_t = x_0'(S - M^*)x_0 \tag{14}$$

Again from (12c) it can be seen that these costs are certainly nonnegative and unless stabilization is perfect will in fact be positive.

We now consider the case where there are zero control costs so that $N=0$. As long as the dynamic equation (6b) includes some lags in u_t, the corresponding matrix M^* will include some blocks of N which are zero (see (9)), and consequently M^* can only be positive semidefinite. In this case the optimal policy simplifies to:[15]

$$u_t = Kx_{t-1} \tag{12a}$$

where

$$K = -(B'SB)^{-1}B'SA \tag{12b$'$}$$

and

$$S = M^* + (A + BK)'S(A + BK) \tag{12c}$$

Two things should be noted about the solution. First, it will exist only provided $(B'SB)$ is nonsingular. As long as equation (6b) includes some lags in u_t then this requirement need not impose any constraints on the magnitudes of q_1 and q_2.[16] Secondly, since M^* is now only positive semidefinite, S can also now only be shown to be positive semidefinite. Consequently, the quantity $V_t \equiv x_t' Sx_t$ is no longer strictly positive and *cannot* serve as a Liapunov function for the entire vector x_t. Thus the proof adopted above to establish stability is not applicable in the present case.

Nevertheless we are still able to deduce some stability properties of the system. Since the cost matrices M corresponding to the target variables y_{t-i} are strictly positive definite, the submatrix M_1^* (see (9)) is also strictly positive definite. If we partition S so that the submatrix S_1 corresponds to M_1^*, it follows from (12c$'$) that S_1 is in fact strictly positive definite, so that the quantity

$$V_{1t} = x_{1t}'S_1x_{1t}$$

where $x_{1t} = [y_t, \ldots, y_{t-(m-1)}]'$, is a positive function. Since it can be readily shown that $V_{1t} < V_{1,t-1}$, it follows that x_{1t} and its subvectors y_{t-i} tend to zero, so that target stability is still established. Note the fact that we are unable to establish instrument stability does *not* necessarily imply instability, since the Liapunov method yields sufficient but not necessary stability conditions. However, Holbrook's example, which we shall reconsider in this framework in Section III, indicates how instrument instability may in fact arise.

Even with zero control costs, instrument stability can still be established in the important case where the dynamics of the system can be described by an equation of the form

$$y_t = A_1y_{t-1} + \ldots + A_my_{t-m} + B_0u_t \tag{15}$$

In this case the cumulated effects of past policies on current values of the target variables are summarized by the lagged values of the targets. It includes as a special case what one might call a "generalized Koyck distributed lag," namely,

$$y_t = A_1y_{t-1} + B_0u_t \tag{15$'$}$$

so that $A = A_1$, $B = B_0$. However it also includes many much more general and flexible lag patterns as well.

With $n=0$, M^* does not contain any block of N so that M^* is strictly positive definite if, and only if, M is strictly positive definite, as we assume. Furthermore, as pointed out in footnote 11, the strict positive definiteness of M^* is sufficient to ensure that the solution S also has the property. Hence, once again $V_t = x_t' S x_t$ may serve as a Liapunov function and repeating the earlier argument we establish the stability of x_t. Now, however, x_t consists only of $y_t, \dots, y_{t-(m-1)}$ and thus convergence of x_t implies that each $y_t \rightarrow 0$; i.e., target stability. Moreover from the linear feedback property of the optimality condition (12) we see that this in turn implies $u_t \rightarrow 0$, thereby ensuring instrument stability as well.[17]

---

[15] These remarks concerning the positive semidefiniteness of M^* also apply in the case where N is only positive semidefinite. However, in that case N would still appear in the optimal policy (12).

[16] Recalling the dimensionality of B and S from (7) and (12c), respectively, a necessary (but not sufficient) condition for $(B'SB)^{-1}$ to exist is that $mq_1 + nq_2 \geqq q_2$. As long as $n \geqq 1$, so that at least one lag in u_t is included in (6b), this condition is automatically assured. Restrictions are imposed only if $n=0$.

A necessary condition for the existence of the inverse matrix $(B'SB)^{-1}$ is that $mq_1 \geqq q_2$, which for the Koyck lag case $(m=1)$ simplifies to the requirement that the number of instruments cannot exceed the number of targets. If we consider the case treated by Holbrook where $q_1 = q_2$ and B_0 is nonsingular, we obtain

$$(16) \quad K = -(B'SB)^{-1}B'SA = -B^{-1}A$$

so that the optimal policy simplifies to[18]

$$u_t = -B_0^{-1}[A_1 y_{t-1} + A_2 y_{t-2} + \dots + A_m y_{t-m}]$$

Substituting for u_t in (15) we immediately see further that $y_t = 0$, for all t, so that in this case exact stabilization in each period is possible. Note, further, that equation (15), together with (12c′) and (14) implies that the total cost in this case is zero, as one would indeed expect.

To summarize the deterministic case, if the policy maker ascribes costs to all targets and instruments, then the optimal policy implies stability in both targets and instruments. If, as is frequently assumed, control is costless then the possibility of instrument instability may indeed arise although target stability will still prevail.[19] In this case one might quite legitimately ask why worry about instrument instability if there are no costs associated with having instruments away from their desired values? Leaving that issue aside, even if there are zero costs to control, instrument instability can be ruled out provided the dynamics can be described by an equation such as (15). That is, instrument stability is still assured if the past effects of policy variables are summarized in past values of the target vector y_t. As discussed this includes the well-known Koyck lag as a special case, but it also includes many other more general lag structures as well.

II. Stability in a Stochastic Model

It is clear from the previous section that the strict positive definiteness of S plays a crucial role in ensuring the stability of both targets and instruments. In this section we demonstrate the analogous result for the case where the parameters of A and B are stochastic, and the dynamic equation includes a vector of additive disturbances. Thus we have

$$(17) \quad x_t = (A + V_t)x_{t-1} + (B + W_t)v_t + e_t$$

where the matrices $V_t = (v_{ijt})$, $W_t = (w_{ijt})$ are matrices whose elements are random variables independently distributed through

---

[17] Writing

$$v_t = \sum_{i=0}^{n} B_i u_{t-i}$$

it follows from this result that for the more general dynamic model (6b) $v_t \rightarrow 0$, so that in the limit, the instrument vector must satisfy the difference equation $B_0 u_t + B_1 u_{t-1} + \dots + B_n u_{t-n} = 0$.

[18] With $n=0$, the matrices A, B defined in (7) simplify to

$$A = \begin{bmatrix} A_1 & A_2 \dots & A_m \\ I & & \cdot \\ \cdot & I & \vdots \\ \vdots & & \vdots \\ 0 & \dots\dots & I \end{bmatrix} \quad B = B_0, \text{ respectively.}$$

[19] Presumably, some targets could be unstable if M were only positive semidefinite.

time and e_t is a vector of intertemporally independently distributed random variables. We assume that (i) the elements of V_t are correlated; (ii) the elements of W_t are correlated; (iii) V_t is correlated with W_t; (iv) both V_t and W_t are uncorrelated with the additive disturbance vector e_t; (v) all random variables have zero means and finite variances, assumed to be constant through time.

Because of the additive disturbances, the natural objective function

$$E\left[\sum_{t=1}^{\infty}(x_t' M^* x_t + u_t' N u_t)\right]$$

where E denotes expected value, will diverge and hence the criterion must be modified. Following Wonham, a natural alternative is to minimize the expected cost per period under stationary conditions, i.e., assuming that all random variables have converged to their limiting distributions. This is essentially minimizing the expected long-run average cost. As shown by Wonham the solution to the finite-time horizon stochastic problem that is the minimum of the expression:

$$E\left[\sum_{t=1}^{T}[x_t' M^* x_t + u_t' N u_t]\right]$$

does indeed converge to the optimal control for this stationary problem. Hence the proposed criterion is in fact quite general.

As shown by Wonham, one important difference between the deterministic and stochastic stabilization problem is that in the latter case there may in fact be no optimal policy.[20] Assuming that the objective is to minimize

$$E[x_t' M^* x_t + u_t' N u_t]$$

under stationary conditions, we obtain the following results.[21]

THEOREM 1: *An admissible control exists if (A, B) is stabilizable*[22] *and*

$$(18)\quad inf\left\|\sum_{t=1}^{\infty}(A' + K'B')^t[E(V'V) + E(V'W)K + K'E(W'V) + K'E(W'W)K](A + BK)^t\right\| < 1$$

Basically this theorem asserts that an admissible control exists provided the disturbances in the parameters are not too great. The condition is independent of any additive noise and will certainly be satisfied under deterministic conditions. Thus we can state:

THEOREM 2: *Assuming (A, B) is stabilizable, and (18) is satisfied and M^* and N are positive definite, then the optimal control exists and is of the form*[23]

$$(19a)\quad u_t = Kx_{t-1}$$

where

$$(19b)\quad K = -[N + B'SB + E(W'SW)]^{-1}\cdot[B'SA + E(W'SV)]$$

and S is the positive definite solution to the equation

$$(19c)\quad S = M + K'NK + (A + BK)'S(A + BK) + E[(V + WK)'S(V + WK)]$$

---

[20] For a proof of these theorems in the continuous case, see Wonham. These continuous results have been slightly extended in my 1973 paper, where they have been applied to the derivation of optimal stabilization policies for a multiplier-accelerator model.

[21] The assumption that A, B are constant is not necessary. If A, B, M^*, N are time dependent, the differences in the optimal policy will be similar to those differences which arise in the deterministic case. See fn. 12.

[22] The pair (A, B) describing the dynamic system $x_t = Ax_{t-1} + Bu_t$ is stabilizable if, and only if, there exists a matrix K such that the matrix $A+BK$ is stable (i.e., the real parts of eigen values are less than unity in magnitude). The quantity $\|x\|$, where x is a vector, is defined to be the usual Euclidean norm. For a matrix A, $\|A\|$ is taken to be the Euclidean operator norm and can be defined by $\|A\| = \max[\|Ax\| : \|x\| = 1]$.

[23] The results apply even if the variances are functions of time.

The analogy between (19) and (12) is immediate. As one would expect, the optimal policy under stochastic conditions depends upon the stochastic disturbances in the parameters (but not on the additive disturbances) and that if these are zero, the two solutions coincide.

With stochastic elements present the deterministic notion of stability must be modified, although as we now show analogous stability results are obtained. Making use of results due to R. S. Bucy and Harold Kushner, provided S is strictly positive definite, the function

$$V_t = x_t' S x_t$$

can serve as a stochastic Liapunov function.[24] Thus the vector x_t will converge to zero with probability one and therefore be asymptotically stable with probability one provided

$$E[V_t \mid x_{t-1}] < V_{t-1} \qquad \text{for all t}$$

Asymptotic stability with probability one is extremely strong, being a direct extension of the deterministic definition. It requires the random function to ultimately converge to the nonstochastic equilibrium and obviously cannot be satisfied in the presence of an additive disturbance with constant variance.

Suppose that there are no additive disturbances so that $e_t \equiv 0$. Thus when the optimal policy (19) is applied, x_t is given by:

$$(20) \quad x_t = [(A + BK) + (V + WK)]x_{t-1}$$

and taking conditional expectations we find

$$\begin{aligned} &E[V_t \mid x_{t-1}] - V_{t-1} \\ &\quad = x_{t-1}'[(A + BK)'S(A + BK) \\ &\qquad - E[(V + WK)'S(V + WK)] - S]x_{t-1} \\ &\quad = - x_{t-1}'[M^* + K'NK]x_{t-1} < 0 \end{aligned}$$

implying asymptotic stability with probability one.

In direct analogy to the deterministic case, the sufficient condition for asymptotic stability with probability one of both targets and instruments depends crucially on S being strictly positive definite. As before, the strict positive definiteness of M^* suffices to ensure the same for S and hence guarantees that both targets and instruments are stochastically stable in this very strict sense. Consequently, all the results concerning instrument and target stability obtained under certainty hold analogously with respect to this form of stochastic stability in the case where the coefficients of the system are subject to random disturbances. However, the introduction of these stochastic disturbances does introduce one important difference. The matrix

$$E[(V + WK)'S(V + WK)]$$

is at least positive semidefinite so that if S is only positive semidefinite in a deterministic model (in which case instrument instability may occur), it is possible that it will become *strictly* positive definite with the introduction of the multiplicative stochastic terms. In other words, it is possible that the introduction of such stochastic disturbances may induce instrument stability into the system. Basically the reason for this stems from the fact that the presence of stochastic disturbances causes the policy maker to apply his instruments more cautiously and to reduce the intensity of the adjustment.

Taking into account the additive disturbances, a weaker form of stochastic stability is still found. Suppose that the stochastic Liapunov function defined above has the property $E[V_t \mid x_{t-1}] < V_{t-1}$. Hence the system without additive noise

---

[24] For a detailed discussion of stochastic stability and proofs of the relevant theorems, see Kushner. A summary of the main results, as well as a brief discussion of stochastic Liapunov functions is given in my 1968 paper.

$$(17') \quad x_t = (A + V_t)x_{t-1} + (B + W_t)u_t$$

is stable with probability one and can also be shown to be stable in the mean (see Kats and Krasovskii). Using some results of Kats and Krasovskii, it can be shown that if the variance of e_t is finite, then the additively disturbed system (17) is stable in probability. By this it is meant that x_t remains within some prescribed neighborhood of the equilibrium with some assigned probability. Hence the same conditions that ensure the strict positive definiteness of S ensure stability in probability of the optimally controlled additively disturbed system.

III. An Example

To illustrate the above results and in particular how the introduction of stochastic disturbances in the parameters may eliminate instrument instability, we shall consider the following simple model, which is essentially that studied by Holbrook.

In our notation the system can be represented as:

$$(21) \quad y_t = w_0 u_t + w_1 u_{t-1}$$

where w_0, w_1 are parameters, which for the moment are unrestricted, but which frequently are interpreted as weights and thus are required to sum to unity, say. The policy maker wants to keep y_t as near to its target (taken for convenience to be zero) so that the formal optimization problem is to minimize

$$(22a) \quad \sum_{t=1}^{\infty} y_t^2$$

subject to

$$(22b) \quad \begin{bmatrix} y_t \\ u_t \end{bmatrix} = \begin{bmatrix} 0 & w_1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ u_{t-1} \end{bmatrix} + \begin{bmatrix} w_0 \\ 1 \end{bmatrix} v_t$$

This is written in the form of equation (8) so that the definitions of A, B are clear. Thus solving for the optimal policy we obtain[25]

$$(23) \quad S = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$-(B'SB)^{-1}B'SA = (0, -w_1/w_0)$$

yielding the optimal control policy

$$(24) \quad u_t = -(w_1/w_0)u_{t-1}$$

Substituting into (21) this immediately implies $y_t = 0$, for all t, so that using the optimal policy, the decision maker is able to stabilize perfectly and to incur zero costs.

This of course is the identical solution to that of Holbrook which is as it should be, although his optimization is somewhat implicit. But the most important point is that S is now only positive semidefinite so that using the quadratic Liapunov function we are unable to establish instrument stability. Indeed this is confirmed by equation (24), which as Holbrook pointed out will imply instability if $w_1 > w_0$.[26]

Let us now suppose that the parameters w_0, w_1 are subjected to random disturbances e_0, e_1, assumed to be uncorrelated through time, with zero means and known variances σ_0^2, σ_1^2, respectively, and a contemporaneous correlation coefficient of ρ. For the present we maintain our assumption that the weights w_0, w_1, are unconstrained. Thus in terms of the notation of Section II, the stochastic analogue of (21) is

---

[25] As mentioned in fn. 11, S may not be unique in this case. Actually, for $w_1 > w_0$ we obtained a second matrix S given by

$$S = \begin{pmatrix} 1 & 0 \\ 0 & w_1^2 - w_0^2 \end{pmatrix}$$

This is strictly positive definite, but it can be readily verified that the optimal policy corresponding to it is inferior to (24).

[26] As Holbrook discusses, instrument stability can be achieved by only partially adjusting the instrument. This of course is true, but would not be optimal given our objective function (22a), which is also the objective function which Holboook implicitly has in mind.

$$(25) \qquad \begin{bmatrix} y_t \\ u_t \end{bmatrix} = \begin{bmatrix} 0 & \bar{w}_1 + e_1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ u_{t-1} \end{bmatrix} + \begin{bmatrix} \bar{w}_0 + e_0 \\ 1 \end{bmatrix} v_t$$

where $\bar{w}_0$, $\bar{w}_1$ denote the means of w_0, w_1, respectively. For simplicity we shall assume the additive disturbances to be identically zero since they do not alter the optimal policy. In the notation of Section II, we can calculate the following expressions required for (19)

$$E(V'SV) = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_1^2 s_{11} \end{bmatrix}$$

$$E(W'SW) = \sigma_0^2 s_{11}$$

$$E(V'SW) = \begin{bmatrix} 0 \\ \sigma_0\sigma_1\rho s_{11} \end{bmatrix}$$

Thus solving equation (19c) we find

$$(26) \qquad S = \begin{pmatrix} 1 & 0 \\ 0 & s_{22} \end{pmatrix}$$

where s_{22} is obtained by solving for the positive root of the quadratic equation:

$$(27) \quad s_{22}^2 + s_{22}[(\bar{w}_0^2 + \sigma_0^2) - (\bar{w}_1^2 + \sigma_1^2)] - [\bar{w}_1^2\sigma_0^2 + \bar{w}_0^2\sigma_1^2 + \sigma_1^2\sigma_2^2 - 2\bar{w}_0\bar{w}_1\sigma_0\sigma_1\rho - \sigma_0^2\sigma_1^2\rho^2] = 0$$

Since $-1 \leqq \rho \leqq 1$, it is clear that

$$(28) \quad \bar{w}_1^2\sigma_0^2 + \bar{w}_0^2\sigma_1^2 + \sigma_1^2\sigma_2^2 - 2\bar{w}_0\bar{w}_1\sigma_0\sigma_1\rho - \sigma_0^2\sigma_1^2\rho^2 \geqq 0$$

so that the positive root yields a value for $s_{22} \geqq 0$.

Thus an admissible solution exists irrespective of the intensity of the disturbances in the stochastic parameters w_0, w_1, so that condition (18) is always satisfied. Furthermore, provided σ_0^2 or σ_1^2 are greater than zero and $-1 \leqq \rho < 1$, then inequality (28) holds strictly in which case it immediately follows that s_{22} is strictly positive. Hence in general, the matrix S defined by (26) is strictly positive definite, although as one would expect, one case in which it does coincide with (23) is in the absence of noise when $\sigma_0^2 = \sigma_1^2 = 0$.

Consequently, the strict positive definiteness of S implies that $x_t' S x_t$ may serve as a stochastic Liapunov function and following the argument of Section II, this implies that the state vector

$$\begin{bmatrix} y_t \\ u_t \end{bmatrix}$$

is asymptotically stable with probability one, implying *both* instrument and target stability. Taking this solution for S, the optimal policy is given by:

$$(29) \qquad u_t = -\frac{[\bar{w}_1\bar{w}_0 + \sigma_0\sigma_1\rho]u_{t-1}}{(\bar{w}_0^2 + \sigma_0^2) + s_{22}}$$

which of course reduces to (24) when $\sigma_0 = \sigma_1 = 0$. Substituting for this optimal policy into (25) we can readily show that this in turn implies $E(y_t \mid y_{t-1}) \neq 0$. Consequently, it is not optimal for the policy maker to attempt to drive the target to equilibrium each period, as he would do under certainty.

The one instance where instrument instability may arise is if $\rho = 1$ so that w_0 and w_1 are perfectly correlated. In this case, the left-hand side of (27) will reduce to zero if

$$(30) \qquad \sigma_0/\bar{w}_0 = \sigma_1/\bar{w}_1$$

so that, not only are the coefficients perfectly correlated, but also they have the same coefficient of variation. Under these conditions $s_{22} = 0$, rendering S only positive semidefinite, in which case problems of instability may arise. In fact substituting (30) into (29) we find $u_t = -(\bar{w}_1/\bar{w}_0)u_{t-1}$, which is of course identical to (24) derived under certainty.

Finally, we should also consider the case where the stochastic parameters are subjected to

$$w_0 + w_1 = 1 \tag{31}$$

since this is a constraint frequently imposed by economists in analyzing lag responses. This requirement simplifies the solution since it implies $\sigma_0^2 = \sigma_1^2 = \sigma^2$ say, and $\rho = -1$, so that we obtain

$$s_{22} = \frac{-(\bar{w}_0 - \bar{w}_1) + [(\bar{w}_0 - \bar{w}_1)^2 + 4\sigma^2]^{\frac{1}{2}}}{2} > 0 \tag{32}$$

Thus as long as the parameters are subject to some noise, then the optimal policy will always imply both target instrument stability with probability one.

To summarize this example we may note the following. If the sum of the stochastic parameters is not constrained, then, provided they are not perfectly correlated, the optimal stochastic stabilization policy will ensure asymptotic stability with probability one of *both* targets and instruments, even though the analogous policy applied under certainty may quite plausibly yield instrument instability. If $\rho = 1$, then complete stability will still be obtained proved the coefficients of variations of w_0 and w_1 are not the same. If, as is frequently assumed, the sum of the weights adds to unity, then stochastic stability of both targets and instruments will always result. Finally, we may note that if additive disturbances are also included in (25), then provided they are of finite variance, the same results apply with respect to the stability in probability of targets and instruments.

The result emerging from this example, namely that stochastic elements have a stabilizing influence, is rather intriguing in view of the fact that in purely descriptive dynamic models such disturbances tend to be destabilizing (see, for example, my 1968 paper and my paper coauthored with Weintraub). Basically the reason for this is that the existence of stochastic elements exerts an inhibiting influence on the policy maker, causing him to adjust his instruments more cautiously than he otherwise would, thereby dampening the adjustment in u_t and making it stable. At the same time it means that typically there is correspondingly less response in the target so that as in this example it may no longer be optimal to drive it to its desired equilibrium each period. In this sense the stochastic disturbances exert a destabilizing effect on y_t, although of course the target still remains stochastically stable. Thus a kind of tradeoff between the two kinds of stability is seen to occur, although overall the entire system is stochastically stable.

REFERENCES

K. J. Astrom, *Introduction to Stochastic Control Theory*, New York 1970.

M. Athans and P. L. Falb, *Optimal Control*, New York 1966.

R. S. Bucy, "Stability and Positive Supermartingales," *J. Differential Equations*, Apr. 1965, *1*, 151–55.

C. G. Chow, "Optimal Stochastic Control of Linear Economic Systems," *J. Money, Credit, Banking*, Aug. 1970, *3*, 291–302.

———, "Optimal Control of Linear Econometric Systems with Finite Time Horizon," *Int. Econ. Rev.*, Feb. 1972, *13*, 16–25.

———, "Problems of Economic Policy From the Viewpoint of Optimal Control," *Amer. Econ. Rev.*, Dec. 1973, *63*, 825–37.

E. M. Gramlich, "The Usefulness of Monetary and Fiscal Policy as Discretionary Stabilization Tools," *J. Money, Credit, Banking*, May 1971, *3*, 506–32.

R. S. Holbrook, "Optimal Economic Policy and the Problem of Instrument Instability," *Amer. Econ. Rev.*, Mar. 1972, *62*, 57–65.

I. I. Kats and N. N. Krasovskii, "On the Stability of Systems with Random Parameters," *PMM J. Applied Mathematics, Mechanics*, 1960, *24*, 1225–45.

H. T. Kushner, *Stochastic Stability and Control*, New York 1967.

J. La Salle and S. Lefschetz, *Stability by Liapunov's Direct Method with Applications*, New York 1961.

S. J. Turnovsky, "Stochastic Stability of Short-Run Market Equilibrium Under Variations in Supply," *Quart. J. Econ.*, Nov. 1968, *82*, 666–81.

———, "Optimal Stabilization Policies for Deterministic and Stochastic Linear Economic Systems," *Rev. Econ. Stud.*, Jan. 1973, *40*, 79–95.

——— and E. R. Weintraub, "Stochastic Stability of a General Equilibrium System Under Adaptive Expectations," *Int. Econ. Rev*, Feb. 1971, *12*, 71–86.

W. M. Wonham, "Random Differential Equations in Control Theory," in A. T. Bharucha-Reid, ed., *Probalistic Methods in Applied Mathematics*, Vol. II, New York 1969.

——— and W. F. Cashman, "A Computational Approach to Optimal Control of Stochastic Saturating Systems," *Int. J. Control*, 1969, *10*, 77–98.

Product Liability: The Case of an Asset with Random Life

By JOHN P. BROWN*

When an asset has a random life, the discounted cost of replacement is uncertain and is therefore a risk. This paper is an analysis of the sharing of that risk among three agents: the seller of the asset; the buyer; and an insurer. I analyze three particular methods of sharing the risk among the agents. First, there is a guarantee or warranty of some duration G. The buyer pays the insurer a premium and, if the asset fails before G, the insurer pays the buyer some amount R, otherwise the insurer pays nothing. Second, there is a standard-form guarantee, identical to the guarantee except that it is provided automatically with the asset by the seller and its provisions are fixed. Third, there is the underlying legal regime, which assigns responsibility for any risks in the absence of any contract such as a guarantee which reassigns them. The common law doctrine of *caveat emptor*, "let the buyer beware," assigns all responsibility for failure of the asset to the buyer as soon as he buys it. This doctrine has apparently been eroded in the courts recently and there are advocates of the quite different doctrine of strict liability or *caveat venditor*, "let the seller beware."

Throughout the paper I assume that the probability distribution of failure of the asset is embedded in the asset and cannot be changed by any actions of any of the agents. Thus, we deal only with what James Buchanan terms a pure accident. By limiting the analysis in this way we shall avoid the complications due to interactions between guarantees, maintenance effort, and moral hazard.[1]

The first task in this paper is to find the necessary conditions for optimality of the amount and duration of a guarantee from the point of view of the buyer. Then it is shown that the optimal guarantee is unique. I then consider the polar case of zero transaction costs with insurance available at actuarially fair rates. I then turn to the case where the failure distribution is unknown to the buyer but is known to the seller and insurer. There it will be shown that under certain conditions the buyer need not know the failure distribution of the asset in order to choose his optimal guarantee, and further, he can infer the failure distribution and transaction costs from the premium function.

Next, I briefly analyze standard form guarantees and discuss the conditions for an optimal supplementary guarantee or exculpatory contract, where appropriate. Finally, I discuss the role of the legal regime and compare it to a standard form contract. There I show, contrary to the claim of some authors, that it cannot be said a priori which legal regime is best.

This paper draws on the economics literature both on optimal insurance and on products liability.[2] Kenneth Arrow pio-

---

* Associate professor of economics, Cornell Law School. An earlier version of this paper was presented at the 1971 North American Summer Meeting of the Econometric Society in Boulder, Colorado, Aug. 1971. I have benefitted from the comments of James A. Hanson, Robert Jones, and Richard B. Westin.

[1] These issues will be dealt with in a later paper.

[2] The economic analysis of what lawyers call products liability is almost entirely contained in a very remarkable and valuable symposium, "Products Liability:

neered the theory of optimal insurance in the classic appendix to his 1963 article, "Uncertainty and the Welfare Economics of Medical Care." Arrow dealt with insurance against a risk of unknown size at a given date. In a more recent article, Vernon Smith developed conditions for optimal casualty and liability insurance, again at a given date. In the present paper the risk is of a given size but an unknown date.

I. The Assumptions, The Notation, and the Model

The Buyer

The buyer of the durable asset is presumed to have a von Neumann-Morgenstern utility function displaying risk aversion U, where $U'>0$ and $U''\leq 0$. Further, it is presumed that the buyer's utility is additive over time and discounted at a constant rate ρ. For mathematical convenience I also assume that the buyer (or his family) will live forever. Thus if $y(t)$ is a stream of income over time, the buyer evaluates it by

$$(1) \qquad \int_0^\infty e^{-\rho t} U[y(t)]dt$$

When that stream is uncertain, the utility at each date is replaced by the expected utility at that date:

$$(2) \qquad \int_0^\infty e^{-\rho t} EU[y(t)]dt$$

In this paper I shall be concerned with minimizing losses, so it will be convenient to define the utility loss function L corresponding to U. Call the normal level of consumption Ω, which we assume is fixed over time. Then for a loss of X dollars, we define the utility loss function as follows:

$$(3) \quad L(X) = U(\Omega) - U(\Omega - X), \quad \text{for } X \geq 0$$

Then the loss function has the following properties:

$$(4) \qquad L(0) = 0$$

$$L'(X) = U'(\Omega - X) > 0$$

and

$$L''(X) = -U''(\Omega - X) > 0$$

Minimizing utility loss is equivalent to maximizing utility.

The Seller

The seller is presumed to be risk averse or, in the limit, risk neutral.

The Insurer

The insurer is presumed to be risk neutral ($U''=0$). That is, he is presumed to be indifferent to exchanging risky gambles for their certainty monetary equivalents.

The Asset

The asset is extremely simple. Like a light bulb, it has only two states, operating and failed. The date the asset will fail is unknown but subject to a known probability distribution. That probability distribution is described by a probability density function $f(t)$, which gives the probability that the asset will fail at age t. The probability density function is assumed nonincreasing over time, that is, $f'(t) \leq 0$.[3] Denote by $\overline{G}$ the largest t for which $f(t) > 0$, $\overline{G} \leq \infty$; $\overline{G}$ is the greatest possible life of the asset.

The cost of the asset is A dollars. I assume that the asset is a necessity like a refrigerator or washing machine in that

---

Economic Analysis and the Law," published in the *University of Chicago Law Review* for Fall 1970. Important contributions by Roland McKean, James Buchanan, Guido Calabresi and Kenneth Bass, and Robert Dorfman as well as a transcript of a conference of these and other economists and lawyers are all published there. W. David Slawson provides fascinating insight into the legal issues raised by standard form contracts.

[3] A nonincreasing probability density function for failure is a standard assumption in reliability theory. Apparently most work in reliability has found this assumption appropriate as well as useful. See R. E. Barlow and F. Proschan.

when the asset fails, the buyer will immediately replace it.

The Guarantee

When the buyer purchases a new asset he may also buy a guarantee from the insurer. He is free to buy any nonnegative guarantee period G. If the asset fails during the guarantee period (at some date t where $t \leq G$), the insurer will pay the buyer an amount R, the net recovery, at that time. If the asset fails after the guarantee period ($t > G$), the buyer receives nothing from the insurer. The premium $P(R, G)$ is a twice differentiable increasing function of both R and G. When $G = \overline{G}$, we call the guarantee a lifetime guarantee.

To give more structure to the problem, it is necessary to make more assumptions about the premium function. We assume that the premium is a positive multiple $\gamma \geqq 1$ of the expected cost of servicing the guarantee. We consider the possibility of two kinds of transaction costs, the sales cost C_s which must be paid at the time of the sale of the guarantee, and claim adjustment costs C_c which are paid when the asset fails if it fails during the guarantee period. Claim adjustment costs will arise whenever anyone other than the buyer bears part of the risk of failure. That party must be assured that there has in fact been a failure and getting such assurance is costly. The expected discounted cost of servicing the guarantee is

$$(5) \quad C(R, G) = C_s + \int_0^G (R + C_c)e^{-\rho t}f(t)dt$$

so the premium $P(R, G)$ is then

$$(6) \quad P(R, G) = \gamma[C_s + (R + C_c)D(G)] \qquad \text{for } \gamma \geq 1$$

where

$$D(G) = \int_0^G e^{-\rho t}f(t)dt$$

Also let $D = D(\overline{G})$. The expected profit from the guarantee is $(\gamma - 1)C(R, G)$. Note for future reference that the partial derivatives of the premium are

$$(7) \quad \frac{\partial P(R, G)}{\partial R} = P_R(R, G) = \gamma D(G)$$

and

$$(8) \quad \frac{\partial P(R, G)}{\partial G} = P_G(R, G) = \gamma(R + C_c)\, D'(G) = \gamma(R + C_c)e^{-\rho G}f(G)$$

The total discounted cost to the consumer measured in utility, of having to replace the asset at time t after having bought a guarantee of level R and length G, we call $K(R, G, t)$:

$$(9) \quad K(R, G, t) = \begin{cases} L[P(R, G)] + e^{-\rho t}L(A - R) & \text{for } t \leq G \\ L[P(R, G)] + e^{-\rho t}L(A) & \text{for } t > G \end{cases}$$

The *expected* total discounted cost of replacing the asset, measured in utility, is called $K(R, G)$ and is

$$(10) \quad K(R, G) \equiv E[K(R, G, t)] = L[P(R, G)] + L(A - R)\int_0^G e^{-\rho t}f(t)dt + L(A)\int_G^{\overline{G}} e^{-\rho t}f(t)dt$$

The first term is the loss due to paying the premium, the second term is the expected discounted future loss due to failure of the asset *during* the guarantee period, and the third term is the expected discounted future loss due to failure *after* the guarantee period. Then

$$(11) \quad K(R, G) = L[P(R, G)] + L(A - R)D(G) + L(A)(D - D(G))$$

One could now take explicit account of the infinitely repetitive nature of the guarantee decisions and formulate the problem as one of dynamic programming. Starting from date zero, the expected cost of replacing the asset the first time it fails is K. The expected cost of replacing it the second time is again K as seen from date t, so the expected cost of replacing the asset the second time as seen from date zero is $KEe^{-\rho t}$.[4] Similarly, the cost of replacing the third time, as seen from date zero, is $KEe^{-\rho 2t}$, and so on. Call $V(R, G)$ the expected discounted cost of the entire future stream of replacements when recovery R and guarantee period G are always chosen. Then we can write $V(R, G)$ as follows:

$$(12)\quad \begin{aligned} V(R, G) &= K(R, G) + K(R, G)Ee^{-\rho t} \\ &\quad + K(R, G)Ee^{-\rho 2t} + \ldots \\ &= K(R, G) + V(R, G)Ee^{-\rho t} \end{aligned}$$

$$(13)\quad V(R, G) = \frac{1}{1 - Ee^{-\rho t}} K(R, G)$$

To find the optimal V, it is sufficient to find the optimal K. Therefore, we may focus our attention on $K(R, G)$ without explicitly considering the dynamic nature of the problem.

II. The Optimal Guarantee

First-Order Conditions for a Minimum

The problem faced by the potential buyer of a guarantee is to evaluate the following expression:

$$(14)\quad \min\left[L(A)D, \min_{R,G} K(R, G)\right] \qquad R \geq 0, G \geq 0$$

That is, he calculates the best guarantee that he can buy, $\min_{R,G} K(R, G)$, and then compares it with the expected loss of not buying the guarantee, $L(A)D$, and chooses the lower.

Of course, most of our analysis will deal with the optimal guarantee, assuming that it is better to buy the guarantee than not. That is, we shall mainly be concerned with the first part of the problem,

$$\min_{R,G} K(R, G) \qquad R \geq 0, G \geq 0$$

where

$$(15)\quad \begin{aligned} &K(R, G) \\ &= L[\gamma(C_s + (R + C_c)D(G))] \\ &\quad + L(A - R)D(G) + L(A)(D - D(G)) \end{aligned}$$

Thus we can restrict our attention to[5]

$$(16)\quad \begin{aligned} K_R(R, G) &= P_R(R, G)L'[P(R, G)] \\ &\quad - L'(A - R)D(G) = 0 \end{aligned}$$

and

$$(17)\quad \begin{aligned} &K_G(R, G) \\ &= P_G(R, G)L'[P(R, G)] \\ &\quad - [L(A) - L(A - R)]D'(G) = 0 \end{aligned}$$

---

[4] $Ee^{-\rho t} = \int_0^{\overline{G}} e^{-\rho t} f(t)dt$. Since the t's are independent and identically distributed, $Ee^{-\rho(t_1+t_2)} = Ee^{-\rho 2t}$, where t_1 and t_2 are the lives of the first and second assets, respectively. Of course the distribution of t is independent of guarantee decisions.

[5] The complete necessary conditions for a minimum when there are nonnegativity constraints are shown by Intriligator, p. 49:

(i)	$K_R(R, G) \geq 0$
(ii)	$K_G(R, G) \geq 0$
(iii)	$RK_R(R, G) = 0$
(iv)	$GK_G(R, G) = 0$
(v)	$R \geq 0$
(vi)	$G \geq 0$

In our problem we can ignore solutions of $R=0$ or $G=0$ because then it will always be better not to buy the guarantee at all. This allows us to ignore conditions (iii) and (iv) as well as the inequalities in conditions (i) and (ii) and equalities in (v) and (vi). If conditions (i) and (ii) are not met with strict equality, then we can be sure that it will be best for the buyer to reject the guarantee and insure himself. In other words, if the best available guarantee is the shortest or smallest, i.e., $R=0$ or $G=0$, the buyer will be better off not to buy any at all. This is true because of the discontinuity in the premium function due to the sales cost. Therefore, we need analyze only those situations where the first two conditions hold with strict equality.

The interpretation of the conditions is straightforward when they are slightly rewritten as follows:

$$P_R(R, G)L'[P(R, G)] = L'(A-R)D(G) \tag{18}$$

and

$$P_G(R, G)L'[P(R, G)] = [L(A) - L(A - R)]D'(G) \tag{19}$$

In both (18) and (19) the marginal cost in utility terms must equal the marginal expected discounted gain in utility terms. In (18), the marginal utility of the increment in the premium must equal the increment in utility of the marginal dollar received on failure of the asset during the guarantee $L'(A-R)$, discounted back to the present and weighted by the probability of failure during the guarantee period $D(G)$. In (19), which deals with changing the length of the guarantee, the marginal utility of the increment in the premium must equal the change in utility due to getting guarantee protection another increment in time $L(A)-L(A-R)$, discounted to the present and multiplied by the probability that the asset will actually fail during the increment $D'(G)=e^{-\rho G}f(G)$.

We display conditions (16) and (17), taking account of our specific assumptions about the nature of the premium, (7) and (8):

$$K_R(R, G) = D(G)[\gamma L'(P(R, G)) - L'(A - R)] = 0 \tag{20}$$

$$K_G(R, G) = D'(G)[\gamma(R + C_c)L'(P(R, G)) - (L(A) - L(A - R))] = 0 \tag{21}$$

Condition (20) can be fulfilled in two ways: either the bracketed term vanishes or $D(G)=0$. But $D(G)=0$ only when $G=0$, a case we are not interested in here. Similarly, condition (21) can be fulfilled in two ways: either when the bracketed expression vanishes or when $D'(G)$ vanishes. This latter can be the case only when $G=\overline{G}$.

We can gain further insight into the structure of the problem when $G\neq\overline{G}$ and $G\neq 0$ by the following simple manipulation. Rearrange (20) and (21):

$$D(G)\gamma L'(P(R, G)) = D(G)L'(A - R) \tag{22}$$

$$D'(G)\gamma(R + C_c)L'(P(R, G)) = D'(G)[L(A) - L(A - R)] \tag{23}$$

Divide (23) by (22), simplify, multiply both sides by $D(G)/D'(G)$, and rearrange terms to get

$$(R + C_c)L'(A - R) = L(A) - L(A - R) \tag{24}$$

Expression (24) has a simple geometric interpretation (see Figure 1). If we go to a point C_c units to the right of the loss function at level $L(A)$ and from there construct a line tangent to the loss function, that point where it is tangent will give us $A-R^*$, hence the optimal value for R, because the tangent satisfies equation (24).

Now to find the optimal G, we divide both sides of (22) by $D(G)$. $D(G)$ exists whenever $G>0$. Rearranging terms, this gives us:

$$L'(P(R, G)) = \frac{L'(A - R)}{\gamma} \tag{25}$$

where $\gamma \geq 1$

Now since L is convex in dollar loss, L' is a monotone increasing function of dollar

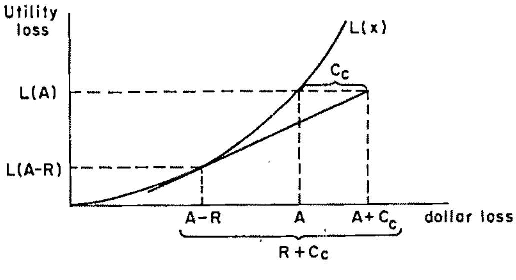


FIGURE 1

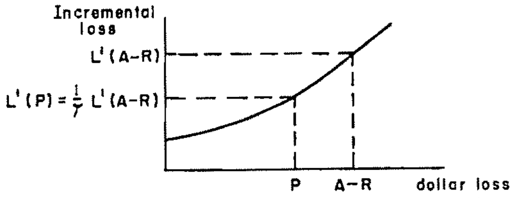


FIGURE 2

loss, which we depict in Figure 2. Therefore, we can invert the L' function to find the optimal premium and then the optimal G.

In this problem the first-order conditions are not only necessary for an optimum but are sufficient as well, because $K(R, G)$ is concave in R and ditonal in G (a function $f(x)$ is said to be ditonal if $-f(x)$ is unimodal). See the Appendix for a proof.

Perfect Competition and Costless Transactions—A Special Case

Perfect competition in the insurance industry implies that there are no profits in insurance so $\gamma = 1$. Costless transactions imply that $C_s = C_c = 0$. Then the premium is

$$(26) \quad P(R, G) = RD(G)$$

and the objective function is

$$(27) \quad K(R, G) = L[RD(G)] + L(A-R)D(G) + L(A)[D - D(G)]$$

with first-order conditions

$$(28) \quad K_R(R, G) = D(G)[L'(RD(G)) - L'(A-R)] = 0$$

and

$$(29) \quad K_G(R, G) = D'(G)[RL'(RD(G)) - [L(A) - L(A-R)] = 0$$

Recall that $D(G) = 0$ only if $G = 0$ and $D'(G) = 0$ only if $G = \overline{G}$. Assume $G > 0$. Then, following the argument leading to (24), if $G < \overline{G}$,

$$(30) \quad RL'(A-R) = L(A) - L(A-R)$$

which can be true only if $R = 0$. This is unlikely, so we can be sure that $G = \overline{G}$. Then (28) becomes

$$(31) \quad L'(RD) = L'(A-R)$$

which implies that

$$(32) \quad RD = A - R$$

and

$$(33) \quad R = \frac{A}{1+D}$$

To recapitulate, the optimal guarantee for perfect competition and costless transactions is $R = A/(1+D)$ and $G = \overline{G}$.

This is a remarkable result. One might have thought that the optimal guarantee under these circumstances would be a lifetime, money-back guarantee; that is, $R = A$ and $G = \overline{G}$. To further convince ourselves that the guarantee we have calculated is in fact optimal, we display and compare the optimal guarantee as calculated, the lifetime, money-back guarantee, and, finally, no guarantee.

$$K\left(\frac{A}{1+D}, \overline{G}\right) = (1+D)L\left(\frac{AD}{1+D}\right) \quad \text{optimal}$$

$$K(A, \overline{G}) = L(AD) \quad \text{lifetime money-back}$$

$$K(0, 0) = L(A)D \quad \text{no guarantee}$$

To compare these we take advantage of the fact that for convex functions such as L,

$$L(\alpha X) < \alpha L(X) \qquad \text{for } \alpha < 1$$

To show that $K(A, \overline{G}) < K(0, 0)$, let $\alpha = D$ and $X = A$ to get

$$(34) \quad L(AD) < DL(A)$$

To show that $K(A/(1+D), \overline{G})$ is in fact superior to $K(A, \overline{G})$ let $\alpha = 1/(1+D)$ and $X = AD$. Then

$$L\left(\frac{AD}{1+D}\right) < \frac{1}{1+D} L(AD)$$

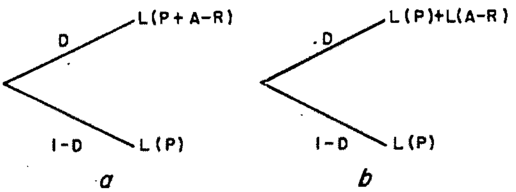


FIGURE 3

and multiplying both sides by $1+D$ we get our result:

$$(35)\qquad (1+D)L\left(\frac{AD}{1+D}\right) < L(AD)$$

It is instructive to consider why a lifetime money-back guarantee is not optimal. Consider the simple lottery of losing A with probability D and losing nothing with probability $(1-D)$. Its expected utility is

$$DL(A) + (1-D)L(0) = DL(A)$$

If one can reduce the loss by R on payment of a premium of $P(R)$, then the gamble's expected utility becomes

$$DL(A + P(R) - R) + (1-D)L(P(R))$$

which is shown in Figure 3a. The optimal guarantee becomes $R=A$ if the premium is actuarial. However, this simple gamble is not the analogue of the intertemporal gamble we have been discussing. For us the appropriate gamble is

$$D[L(A-R) + L(P(R))] + (1-D)L(P(R))$$

which is depicted in Figure 3b. Here the optimal guarantee is $R=A/(1+D)$ as in the intertemporal problem. The difference comes about because the premium is paid at a different time from the cost of replacement, so that the loss is not calculated by adding the premium and replacement cost together. This is then a less severe loss than if the premium and the costs were added together; hence, less insurance is necessary.

III. The Ignorant Buyer

It is an interesting characteristic of the problem as we have posed it that if the buyer knows transaction costs and the insurer's markup (γ), he does not need to know the failure distribution of the asset. He can calculate the optimal R and G without knowing $D(G)$ because both $D(G)$ and $D'(G)$ drop out in the calculation of the optimal conditions between equations (23) and (24). All that the buyer needs to know is his own loss function and the cost and markup parameters.

In fact, the premium function $P(R, G)$ is sufficient information for him to infer the failure probability density function $f(t)$, and the cost and markup parameters. It is necessary for this purpose that the buyer only know the form of the premium function and ρ, the insurer's discount parameter. First, to find $\gamma(1+C_c)$, integrate as follows:

$$(36)\qquad \int_0^{\bar{G}} e^{\rho t} P_G(1, t)dt = \int_0^{\bar{G}} \gamma(1+C_c)f(t)dt$$

$$= \gamma(1+C_c)$$

Then, knowing $\gamma(1+C_c)$, we can calculate $f(t)$ from (8):

$$(37)\qquad f(t) = \frac{e^{\rho t}}{\gamma} P(1, t)$$

To calculate C_c proceed as follows. Consider two values of R, for example, 1 and 2. For some value of G let

$$(38)\qquad X = \frac{P_G(1, G)}{P_G(2, G)}$$

By (8) we know

$$X = \frac{\gamma(1+C_c)D'(G)}{\gamma(2+C_c)D'(G)}$$

$$= \frac{1+C_c}{2+C_c}$$

so

$$(39)\qquad C_c = \frac{2X-1}{1-X}$$

We can now calculate γ since we know $\gamma(1+C_c)$. Knowing γ, we can calculate C_s since $P(0, 0)=\gamma C_s$. Therefore, we can conclude that knowledge of the form of the premium function plus the value of the premium function allows one to infer the parameters of the premium function plus the failure density function.

Typically, however, the available premium function is not so complete and not so informative. How does the buyer of the guarantee who is ignorant of the failure distribution proceed? From the theory of optimal statistical decisions we know that the buyer should quantify his vague notions about the failure distribution in the form of a probability distribution and use it as a substitute for the unknown true distribution. This approach, known as Bayesian, is thoroughly covered by Howard Raiffa and Morris DeGroot.

Assume that the buyer knows with certainty the form of the failure distribution but not the parameters. The seller and the insurer still know the complete distribution with certainty. To be more concrete, let us assume that the true failure distribution is exponential with intensity parameter μ, so the probability density function is $f(\mathrm{t})=\mu e^{-\mu \mathrm{t}}$.[6] The buyer is then presumed to know that the form of the distribution is exponential, but he does not know the value of μ. Instead, the buyer's beliefs about the true value of μ can be expressed as a probability distribution about the true value of μ with density $\gamma(\mu|r, s)$. An interpretation for the parameters r and s, which completely describes his beliefs about the failure distribution, is that he acts as if he had observed r failures in s years. A convenient and desirable property of this natural conjugate distribution, $\gamma(\mu|r, s)$, is that if the buyer with beliefs described by r and s buys an asset and observes it fail after t years, his new beliefs about the asset will be described by a distribution in the same family with parameters $(r+1, s+\mathrm{t})$.

The natural conjugate distribution for the exponential distribution is the "*gamma*-1" (Raiffa and Schlaifer) with density

$$(40) \qquad \gamma(\mu \mid r, s) = \frac{e^{-\mu s}(\mu s)^{r-1}s}{(r-1)!} \quad \text{for } \mu > 0, \; r = 1, 2, \ldots, \; s > 0$$

The prior density of t is now

$$(41) \qquad f(\mathrm{t} \mid r, s) = \int_0^\infty f(\mathrm{t} \mid \mu)\gamma(\mu \mid r, s)d\mu = \frac{s^r}{(s+\mathrm{t})^{r+1}}$$

The buyer will use $f(\mathrm{t}|r, s)$ as the failure density function in his calculations to find the optimal guarantee outlined in Section II above.

It is clear that if experience is the only source of information and the asset is relatively long-lived, as a refrigerator or automobile, it will be a very long time before the buyer can build up sufficient experience to be well informed. In this case his initial priors are based not on experience but on other sources of information such as advertising. Here it would be particularly valuable to have at least flexible standard form guarantees so that for different fees he could buy different length guarantees. The information inherent in the premiums charged for the different lengths of guarantee would allow the buyer to make better inferences about the failure distribution of the asset. It would seem to be an altogether appropriate use of governmental authority to require manufacturers of assets, if they provide guarantees at all, to provide guarantees of several significantly different lengths at different prices.

---

[6] See Bennett Fox for a similar treatment of the more general Weibull distribution.

IV. Standard Form Contracts

Now, of course, the vast preponderance of warranties are not bought separately from a third-party insurer and carefully chosen from an infinite variety of possible warranty contracts, as our model would suggest. Instead, they are standard form contracts which are provided with the asset by the seller. Typically they are not even read or understood by one party to the contract, and often they cannot be read until after the transaction is completed. Examples of the latter are insurance contracts which are usually not received until after the insurance is bought and product warranties which are placed inside packages and which cannot be seen until the package is opened.

The model we have constructed above, even though it seems quite remote from the standard contracts mentioned here, is still quite useful in this context, for certainly the criterion which ought to be used in judging the provisions of standard form warranties should be the choice of the buyer if he were free to choose. The optimal guarantee described above ought to be the standard by which the standard form is judged.

The reasons for standard form contracts are clear. They economize on sales costs. Effectively, the seller acts as agent for the buyer and buys insurance for him. The costs of the transaction with the insurer can then be spread over all the buyers of the asset. If the number of buyers is sufficiently large, the sales cost of the guarantee is effectively zero, that is, $C_s = 0$. Since adjustment costs C_c are costs conditional on the occurrence of failure and are the costs of transactions between the buyer and the insurer, they are unaffected by whether the buyer or his agent buys the guarantee.

The price of the asset now reflects the cost of the guarantee as well as the cost of production of the asset itself. If A_f is the cost of the asset with a standard form guarantee and R_f and G_f are the parameters of the guarantee,

$$\begin{aligned} A_f &= A + \gamma(R_f + C_c)D(G_f) \\ &= A + P(R_f, G_f) - \gamma C_s \end{aligned}$$

Has the cost of the asset gone up? No. The seller now presents the buyer with a tie-in sale of an asset plus an insurance contract, the price of which is lower than the two bought separately. If the standard form guarantee is the optimal guarantee for the buyer, little more can be said.

If the standard form guarantee is less than the buyer would have had, the buyer can enter into a supplementary guarantee, and the total cost would be no greater than if he had bought the entire guarantee directly from the insurer (if the sales cost for a supplemental guarantee is the same as for a regular guarantee).

It is a trivial extension of our analysis to find the optimal supplementary guarantee. What changes is the limits of integration (from 0 to G_f) for the choice of G and the amount of uninsured loss (from A to $A - R_f$) for the choice of R.

But what of the case where the standard form guarantee is too extensive rather than too little? Then the buyer could sell an exculpatory clause which limits the liability of the warrantor to the seller, or equivalently, he could reinsure the insurer. In that case, the buyer would "insure" the seller, who is responsible for all failures until G_f, against any failures between G^* and G_f. In that way the buyer effectively reduces his own insurance. Presumably it is costly to enter into such an exculpatory transaction; call the sales cost C_{s^*}. Again it is a trivial extension of the work in Section II to calculate the optimal exculpatory clause, so we shall not do it here.

The buyer will choose to buy a supplementary guarantee if the net gain in utility terms is greater than the transaction costs he incurs, or he will provide an exculpatory clause if the benefit exceeds the transaction costs. Essentially the buyer cannot lose if the sales costs of the supplementary guarantee and of the exculpatory clause are no higher than for a standard guarantee, and the buyer will gain by having the standard form guarantee be "close" to his own optimal guarantee. How close is close depends on the magnitude of the transaction costs.

It is an interesting question, though beyond the scope of this paper, what array of standard form contracts will be offered by a competitive market, by a monopoly, and by competition among a few. It would seem that the problem is analogous to the problem of spatial competition, as discussed by Harold Hotelling, among others.

V. *Caveat* Whom?

By now we have developed an extensive framework which is useful in discussing the differences among legal regimes. The legal regime assigns liability in the first instance before it is reassigned by contract. Recently there has been a gradual shift in the law away from the traditional regime *caveat emptor*, "let the buyer beware," toward a regime which is sometimes called strict liability, or *caveat venditor*, "let the seller beware." Implicitly our discussion so far has been assuming a legal arrangement of *caveat emptor;* the seller has had no responsibility for failure at any date. Let G_L be the date at which the legal responsibility for failure of the asset turns from the seller to the buyer. Put this way, we have assumed that $G_L=0$ in the above analysis, hence *caveat emptor. Caveat venditor* would give us $G_L=\overline{G}$, that is, the seller would be responsible throughout the lifetime of the guarantee. Of course, there is the possibility of intermediate arrangements with intermediate values of G_L.[7]

In any of these possible regimes the parties are free to reallocate the liability by mutually agreeable contracts. In the case of *caveat emptor*, the buyer, who is the holder of the liability in the first resort, is free to buy a guarantee from the seller or from an insurer. Earlier sections of this paper have dealt with the optimal guarantee for the buyer to choose to buy. In the case of *caveat venditor*, the seller, who is the holder of the liability in the first resort, is free to buy insurance from a third party insurer or transfer the liability back to the buyer by buying an exculpatory agreement from the buyer, which relieves the seller of liability for failure of the asset. An exculpatory agreement is effectively a way of the buyer insuring the seller and it has the advantage over explicit insurance of not requiring any transactions when there is a failure, hence, no claim adjustment costs.

Note that there is complete symmetry between *caveat emptor* and *caveat venditor*. All the rights and duties which were the buyer's when there is a failure become the rights and duties of the seller. Other writers, such as McKean and Buchanan, implicity assumed that *caveat venditor* also implied that sale of exculpatory clauses was illegal, so that they were discussing the case of what is here called *caveat venditor* plus compulsory insurance. Then Buchanan's surprising result, that *caveat emptor* is Pareto-preferable to *caveat venditor*, is in fact a result that insurance should not be compulsory, which is not surprising.

A second aspect of the symmetry between the two *caveats*, noted by Guido Calabresi, is that just as under *caveat emptor*, the buyer is free to choose the seller with the safest product, so under *caveat venditor* the seller should be free to discriminate among buyers according to how risky he expects them to be. This kind of symmetry was also assumed away by Buchanan.

---

[7] We are ignoring the other dimension of the liability —how much should the responsible party pay upon failure, i.e., R. We do this to keep the discussion in the simplest possible framework.

In our model the role of the legal regime is identical to that of a standard form contract. It determines the date (and amount) at which the responsibility of the seller ends, without further contracting. Of course, this further contracting may be a standard form contract. Thus, the analysis of the previous section applies to the problem of legal regimes as well. When the legal regime is other than *caveat emptor*, the price of the asset will have to be higher to reflect the cost of insuring the asset. If the law prescribes R_L and G_L, then the price charged by the seller will be $A_L = A + P(R_L, G_L)$. Has the cost of the asset gone up because of the change in the legal regime? No. Just as in the case of a standard form guarantee, the product offered by the seller has been changed. There is a tie-in sale with the result that some guarantees are less expensive than before and some are more expensive. Those buyers who wanted a guarantee close to the guarantee offered in the first instance by the law are made better off by the change, and those buyers who preferred the guarantee available in the first instance before the law was changed are made worse off.

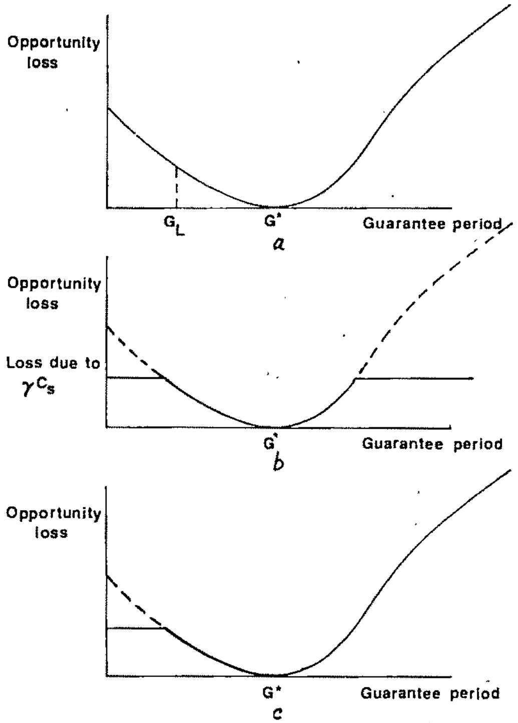


FIGURE 4

Figure 4a makes this more clear. There we have drawn the opportunity loss due to being forced to buy a nonoptimal guarantee of length G, e.g., G_L. For simplicity consider R as given at $\overline{R}$. When the optimal G is positive and finite, Figure 4a is simply $K(\overline{R}, G) - K(\overline{R}, G^*)$. Figure 4a shows the loss if the buyer does not contract for a different guarantee. On payment of an additional γC_s the buyer can convert to the optimal guarantee, so that puts an upper bound on the loss due to the wrong G_L, as depicted in Figure 4b.

The loss due to an inappropriate legal regime decreases as C_s decreases until there is no cost to having the inappropriate legal regime. This is a form of the famous Coase theorem applied to our problem.[8] Notice that the Coase theorem states that when transaction costs are nil, one is indifferent to the legal regime. Here we have the somewhat sharper result that if only sales costs for guarantees are nil, the Coase theorem holds. It is interesting to note that when there are sufficiently many different standard form contracts or if standard form contracts are sufficiently flexible, there are a sufficient number of values of G where sales costs are zero and the Coase theorem holds.

Figure 4c shows the loss due to an inoptimal legal regime when exculpatory agreements are illegal, as in Buchanan's example.

In every situation where C_s is not nil, the choice of the optimal legal regime from society's point of view requires in principle the calculation of the distribution of buyers over optimal G and comparing the losses with the gains. We cannot say, a

---

[8] See Ronald Coase.

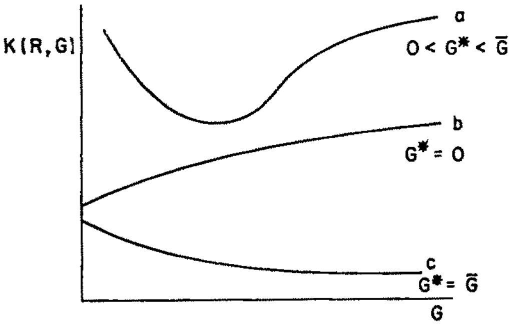


FIGURE 5

priori, what the optimal legal regime should be.

APPENDIX

To show that $K(R, G)$ is convex in R we need only evaluate its second derivative in R:

$$\text{(A1)}\quad K_{RR}(R, G) = D(G)[\gamma L''(P(R, G))\gamma D(G) + L''(A - R)] \geq 0$$

$K(R, G)$ is not uniformly concave in G, that is, $K_{GG}(R, G)$ is not always nonnegative. However, we can show that $K(R, G)$ is ditonal in G, which is sufficient for our purposes. That is,

$$\text{(A2)}\quad K(R, G_1) > K(R, G_2) \quad \text{for } G_1 < G_2 \leq G^*$$

and

$$\text{(A3)}\quad K(R, G_1) < K(R, G_2) \quad \text{for } G^* < G_1 < G_2$$

To demonstrate (A2), it is sufficient to show that whenever $K_G(R, G) \leq 0$, $K_G(R, G-\delta) < 0$ for $0 < \delta < G$. As we see from (21), $K_G(R, G) \leq 0$ if, and when G is finite, only if

$$\text{(A4)}\quad \gamma(R + C_c)L'(P(R, G)) \leq L(A) - L(A - R)$$

Now $P(R, G-\delta) < P(R, G)$, so $L'(P(R, G-\delta)) < L'(P(R, G))$. Finally we have

$$\text{(A5)}\quad K_G(R, G - \delta) < K_G(R, G) \leq 0$$

which demonstrates (A2).

To demonstrate (A3) we must show that $K_G(R, G+\delta) > 0$ whenever $G \geq G^*$ and $\delta > 0$. From (21) we see that nonnegativity of $K_G(R, G)$ is assured if

$$\text{(A6)}\quad \gamma(R + C_c)L'(P(R, G)) \geq L(A) - L(A - R)$$

(A6) holds with equality for G^*. For $G^*+\delta$, $P(R, G^*+\delta) > P(R, G^*)$; hence,

$$\text{(A7)}\quad \gamma(R + C_c)L'(P(R, G^* + \delta)) > \gamma(R + C_c)L'(P(R, G^*)) = L(A) - L(A - R)$$

which is our result.

It is a simple extension to show that when the optimal G is 0, $K(R, G)$ is increasing in G throughout, and to show that when the optimal guarantee is a lifetime guarantee, i.e., $G^* = \bar{G}$, $K_G(R, G) < 0$ for finite G. Therefore, there are only three possible shapes for $K(R, G)$ as a function of G. These are shown in Figure 5.

REFERENCES

K. J. Arrow, "Uncertainty and the Welfare Economics of Medical Care," *Amer. Econ. Rev.*, Dec. 1963, *53*, 941–73.

R. E. Barlow and F. Proschan, *Mathematical Theory of Reliability*, New York 1965.

J. Buchanan, "In Defense of *Caveat Emptor*," *Univ. Chicago Law Rev.*, Fall 1970, *38*, 64–73.

G. Calabresi and K. C. Bass III, "Right Approach, Wrong Implications: A Critique of McKean on Products Liability," *Univ. Chicago Law Rev.*, Fall 1970, *38*, 74–91.

R. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, *3*, 1–44.

M. H. DeGroot, *Optimal Statistical Decisions*, New York 1970.

R. Dorfman, "The Economics of Products Liability: A Reaction to McKean," *Univ. Chicago Law Rev.*, Fall 1970, *38*, 92–102.

B. L. Fox, "An Adaptive Age Replacement Policy," Oper. Res. Cen. Working Paper ORC 65-17, Univ. California 1965.

H. Hotelling, "Stability in Competition," *Econ. J.*, Mar. 1929, *39*, 41–57.

M. Intriligator, *Mathematical Optimization and Economic Theory*, Englewood Cliffs 1971.

R. N. McKean, "Products Liability: Trends and Implications," *Univ. Chicago Law Rev.*, Fall 1970, *38*, 3–63.

———, "Products Liability: Implications of Some Changing Property Rights," *Quart. J. Econ.*, Nov. 1970, *84*, 611–26.

H. Raiffa, *Decision Analysis*, Reading 1968.

——— and R. Schlaifer, *Applied Statistical Decision Theory*, Boston 1961.

W. D. Slawson, "Standard Form Contracts and Democratic Control of Lawmaking Power," *Harvard Law Rev.*, Jan. 1971, *84*, 529–66.

V. Smith, "Optimal Insurance Coverage," *J. Polit. Econ.*, Jan. 1968, *76*, 68–77.

Capital Markets and the Dynamics of Growth

By MATHEW SHANE*

Development of a dynamic monetary framework has been a prime objective of monetary theorists since the publication of James Tobin's article "Money and Economic Growth" in 1965. These efforts[1] have generally been limited to investigating the impact of introducing money into a neoclassical aggregate growth model. In this paper, however, I investigate the effects of introducing an equity security into money and growth analysis. This approach has several distinct advantages. In a model which considers only money and capital, saving is merely a decision to invest directly in real capital or money balances, whereas the introduction of a financial market allows us to analyze directly the mechanism by which saving and investment are separated. In terms of developing a policy tool, this approach also provides a basis for directly relating policy actions which are financial in nature to the effects on real variables such as income and capital intensity.

This paper is further distinguished by an explicit consideration of the stock-flow nature of the "monetary" adjustment process and an analysis of global asymptotic stability.[2] There are three potential formulations of the stock-flow relationship. In Tobin's model, both the supply and demand for money, and saving and investment are assumed equal. This may be viewed as an equilibrium-equilibrium stock-flow model. In some analyses of financial markets and investment (see articles by Herschel Grossman, Dwight Jaffee, and Donald Purvis), a stock disequilibrium adjustment hypothesis has been postured. In our context these are disequilibrium-disequilibrium theories, since neither the supply and demand for money nor saving and investment are equal off the growth path. A third stock-flow adjustment hypothesis, however, implies an equilibrium-disequilibrium model, following an approach developed by Horwich (1955, 1964). It is this hypothesis which will be investigated in the present paper.

The formulation of an equilibrium-disequilibrium model offers two clear advantages over other models: 1) Instantaneous stock equilibrium provides a mechanism for establishing prices and interest rates in the short run, while still allowing for a long-run adjustment process generated by excess flow demands. Since the emphasis of this paper is on long-run adjustment, short-run stock equilibrium is a justifiable simplification.[3] 2) Global stability can be investigated without the assumption of instantaneous adjustment. Thus the addition of a security market serves three roles: it acts as the link between real and monetary behavior; it moderates financial disturbances by permitting changes in real variables to occur over an extensive period; and it provides the medium for separating saving and investment decisions.

I. Capital Market Growth Model

The model consists of two markets: a neoclassical production and factor market, and an asset market for real capital, securities and real balances. The flow of funds generated from saving and investment behavior by acting cumulatively on the asset markets changes its composition and thus provides the dynamic elements of the model. Associated with the two markets are two basic sectors, one banking intermediary and one monetary control sector. Households hold securities issued by firms, and money in the form of currency and deposits. Their saving is a flow demand for these assets. Firms hold the entire capital stock, and issue equity securities to finance the accumulation of the capital stock and real balances. The banking intermediary purchases securities as its only asset, and issues demand deposits to both the household and the firm sectors. Lastly, the monetary control sector increases the stock of outside money through open market operations. An explicit presentation of the model follows.

---

* Associate professor, department of agricultural and applied economics, University of Minnesota. An earlier draft of this paper was presented at the Econometric Society Meetings in Detroit, December 28, 1970. I wish to thank George Horwich, Vernon Ruttan, Terry Roe, Wayne Peng, and Hiroshi Ono, for providing comments and suggestions.

[1] See Michael Hadjimichalakis, David Levhari and Don Patinkin, K. Nagatani, Shane, Michael Sidrauski and Duncan Foley, Jerome Stein, F. Takahashi.

[2] However, it was George Horwich (1955, 1964) who first explicitly recognized the stock-flow nature of the macro-monetary mechanism.

[3] Evidence presented by Eugene Fama and others indicates that the security markets adjust very rapidly to new information, and thus the short-run equilibrium assumption about the equity markets is a reasonably good approximation of reality. On the other hand, recent evidence on banking presented by Jaffe indicates that stock disequilibrium might be the usual state of that market.

A. *Production-Factor Market*

It is assumed that a single good $Y(t)$ is produced by a linear homogeneous neoclassical production function of two factors, labor $N(t)$ and capital $K(t)$, and that technological change $A(t)$ is of the labor augmenting type. This relationship is:

$$(1) \qquad y = f(k)$$

where $y = Y(t)/K(t)$ and $k = A(t)N(t)/K(t)$.

Using the production relationship and assuming competition in the factor markets, it follows directly that the real rental of capital (r) and the real wage rate (ω) are:

$$(2) \qquad r = f(k) - kf'(k) = r(k)$$

$$(3) \qquad \omega = Af'(k) = \omega(k, t)$$

Further, the labor force and technological change are assumed to be growing at exogenous rates n and α respectively:

$$(4) \qquad \dot{N}/N = n$$

$$(5) \qquad \dot{A}/A = \alpha$$

Both the capital stock and labor force are assumed given at some initial point (t_0).

B. *Asset Markets*

Household money wealth PW_h consists of money balances M_h, deposits and currency, plus the value of security holdings P_sS_{qh},

$$(6) \quad PW_h = M_h + P_sS_{qh}$$

Firm's assets PW_f consist of capital PK and money balances M_f:[4]

$$(7) \quad PW_f = M_f + PK$$

Combining (6) and (7), another wealth term is derived:

$$(8) \quad PW = M_h + M_f + PK + P_sS_{qh}$$

The money supply M is divided between an outside M^1 and inside M^2 component:

$$(9) \qquad M = M^1 + M^2$$

which is distributed between household (h), business (f) and banking (b) sectors as follows:

$$(9a) \quad M^1 + M^2 = M_h^1 + M_f^1 + M_b^1 + M_h^2 + M_f^2$$

The banking system is assumed to hold securities P_sS_{qb} and currency M_b^1 and to issue deposits M^2,

$$(10) \qquad M^2 = M_b^1 + P_sS_{qb}$$

while the central bank issues outside money M^1 in exchange for securities P_sS_{qc},[5]

$$(11) \qquad M^1 = P_sS_{qc}$$

Equations (9a), (10), and (11) can be used to derive a simplified form of (8). Thus the money holdings of the household and business sectors can be divided between their inside and outside component:

$$(8a) \quad PW = M_h^1 + M_h^2 + M_f^1 + M_f^2 + PK + P_sS_{qh}$$

---

[4] A unique feature of equity securities as distinct from other methods of financing such as bonds is that equities are not a liability of the firm in the traditional sense. This fact is recognized by Steve Taylor in the following statement: "Corporate equities have an ambiguous position in the financial scheme in that they are only residual claims and, in a legal sense at least, not a burden on the issuer" (p. 23).

[5] It is implicitly assumed that neither the banking system nor the central bank has a net worth. The earnings of each of these sectors is redistributed to the nonbanking public sector so as to not effect relative portfolio allocation.

Using (9a) this can be rewritten as:

(8b) $PW = M^1 - M_b^1 + M^2 + PK + P_s S_{qh}$

Lastly, by substituting (10) for M^2 and (11) for M^1, the following relationship results:

(8c) $PW = PK + P_s S_q$

where by definition

(8d) $P_s S_q = P_s S_{qh} + P_s S_{qb} + P_s S_{qc}$

We will now consider the demand for securities in value terms.[6] The household demand for securities $P_s D_{qh}$ is a function of money income, the security interest rate, and wealth. Thus:

$$P_s D_{qh} = D_{qh}(PY, i_s^*, PW) \tag{12}$$

The demand for securities by the banking system $P_s D_{qb}$, their earning asset, is a simple function of outside money:

$$P_s D_{qb} = \theta(i_s^*) M^1 \tag{13}$$

Substituting (11) into (13) yields:

$$P_s D_{qb} = \theta(i_s^*) P_s S_{qc} \tag{13a}$$

It is further assumed that the supply and demand for securities by the control sector is equal:

$$P_s D_{qc} = P_s S_{qc} \tag{14}$$

Summing equations (12), (13a) and (14) give an aggregate demand for securities $P_s D_q$[7]

$$\begin{aligned} P_s D_q &= D_{qh}(PY, i_s^*, PW) \\ &\quad + [1 + \theta(i_s^*)] P_s S_{qc} \\ &= D_q(PY, i^*, PW) \end{aligned} \tag{15}$$

It is assumed that D_q is homogeneous of degree zero in PY and PW. The rate of open market operations is assumed to be exogenous:

$$\frac{\dot{M}^1}{M^1} = \frac{\dot{\overline{P_s S_{qc}}}}{P_s S_{qc}} = u \tag{16}$$

The supply of securities in value terms is positively related to the profits of the business sector $[(r-\delta)PK]$, negatively related to the interest rate on securities i_s^*,

$$P_s S_q = S_q[(r - \delta)PK, i_s^*) \tag{17}$$

and is linear homogeneous in profits. Since Walras' Law holds in the asset markets, the remaining stock equations—the demand for money of the household and business sectors—are determined and the asset equations are complete.

The interest rate on securities i_s^* is determined by instantaneous asset market equilibrium. Thus setting (15) equal to (17),

$$P_s S_q = P_s D_q \tag{18}$$

and solving for i_s^* yields:

$$i_s^* = D_q^{-1}(PY, P_s S_q, PW) \tag{19}$$

Lastly, it is assumed that the factor markets are always in equilibrium. This is consistent with the full employment neoclassical framework. Thus, the short-run behavior of the asset markets is fully determined.

Real saving is a function of real income, the real return on securities and real wealth,

$$S = S(Y, i_s, W) \tag{20}$$

and is linear homogeneous in income and wealth. Real investment is a function of real profits $(r-\delta)K$, the rental of capital r, and the capital stock K,

$$I = I[(r - \delta)K, r, K], \tag{21}$$

and is linear homogeneous in profits and capital. The rate of actual capital formation $\dot{K}/K$ is assumed to be a composite function of the desired rate of investment I/K and the difference between the rate of savings and investment. Thus:

$$\frac{\dot{K}}{K} = \frac{I}{K} + \gamma\left(\frac{S}{K} - \frac{I}{K}\right), \quad 0 < \gamma < 1 \tag{22}$$

---

[6] This section could be developed in terms of the demand for money of the household and business sector. However, since the emphasis of the paper is on the role of security markets, in this case the only financial instrument, this alternative development is preferred.

[7] Equation (13a) implies a simplified money supply function (see Takahaski) and thus the aggregate demand for securities become a composite function of the three nonbusiness sectors.

Although real saving and investment are the central flow equations, additional flow relationships must be specified. Total household saving in nominal terms S^h is equal to the household flow demand for securities S^h_{sec}, deposits S^h_{dep}, and currency S^h_{cur}:

$$S^h = S^h_{sec} + S^h_{dep} + S^h_{cur} \tag{23}$$

At the same time, the business sectors new issues of securities I_{sec} must equal its demand for deposits S^f_{dep}, currency S^f_{cur}, and investment I^n:

$$I_{sec} = S^f_{dep} + S^f_{cur} + I^n \tag{24}$$

By adding (23) and (24) and making the following substitutions,

$$\dot{M}^1 = S^c_{sec} = S^h_{cur} + S^f_{cur} + S^b_{cur} \tag{25}$$

$$S^b_{sec} + S^b_{cur} = S^h_{dep} + S^f_{dep} = I_{dep} \tag{26}$$

where S^c_{sec}, S^b_{sec} are the new demand for securities by the control agent and banking system and I_{dep}, S^b_{cur} are new deposits and reserves in the banking system, we get:

$$S^h + I_{sec} = S^h_{sec} + S^b_{sec} + S^c_{sec} + I^n \tag{27}$$

Rewriting this, where S^d_{sec} is the sum of S^h_{sec}, S^b_{sec}, and S^c_{sec} yields:

$$S^h - I^n = S^d_{sec} - I_{sec} \tag{27a}$$

an excess of nominal saving over nominal investment likewise implies an excess flow demand for securities. This is an important point and demonstrates the interrelationship between the financial and real behavior. Using (27a) we get that the actual rate of security issues $\epsilon = P_s\dot{S}_q/PK$ is equal to the desired rate I_{sec}/PK plus an adjustment coefficient which depends on the excess of nominal saving over nominal investment:[8]

$$\epsilon = \frac{I_{sec}}{PK} + \phi\frac{(S^h - I^n)}{PK} \qquad 0 < \phi < 1 \tag{27b}$$

Letting $s = S^h/PK = S/K$ and $v = I^n/PK = I/K$, we can rewrite (27b) as:

$$\epsilon = \frac{I_{sec}}{PK} + \phi(s - v) \tag{27c}$$

The desired rate of security issues (I_{sec}/PK) must be equal to the desired additions to real balances plus desired investment. This is a flow portfolio decision which is a function of rates of return (r, i_s):

$$\frac{I_{sec}}{PK} = \Phi(r, i_s) + v \tag{27d}$$

where Φ_1, $\Phi_2 < 0$. Substituting (27d) back into (27c) yields:

$$\epsilon = \Phi(r, i_s) + v + \phi(s - v) \tag{27e}$$

With the flow relationships specified, we can now preceed to hypothesize adjustment functions.

The rate of inflation (p) is assumed to be a function of the excess of relative real investment over real saving:

$$p = \lambda_p\left(\frac{I}{K} - \frac{S}{K}\right) \tag{28}$$

where $\lambda p(0) = p^e$ and $\lambda_p' > 0$. The change in the expected rate of inflation $(\dot{p}^e)$ is a function of the difference between the actual and expected rate of inflation, an adaptive expectation hypothesis,

$$\dot{p}^e = \beta(p - p^e), \qquad 0 < \beta < 1 \tag{29}$$

The rate of change in the rental of capital is derived by taking the time differential of equation (2)

$$\dot{r} = kf''(k)k \tag{30}$$

Since $f''(k)$ is negative, this is a positive function of the difference between the effective labor and real capital growth rates.

II. Reduced Form Model

The extensive model presented above is not in a convenient form for analysis. Since we are interested in the long-run properties of the model, a reduced form of the model will be developed in this section. This will involve deriving reduced equations in the fundamental variables capital intensity k, wealth ratio $w=P_sS_q/PK$ and price expectations p^e.

---

[8] Dividing nominal saving and investment by PK gives the real rate of saving and investment. It is clear from the above presentation that households are the only sector which generates net saving.

Dividing (20) by K gives:

$$s=s[f(k),i_s,I+sq[(r-\delta),i_s+p^e]] \tag{31}$$

where $s_1>0$, $s_2<0$, and $s_3<0$.[9] Similarly, from (21), we get:

$$v=v(r-\delta,r,1)=v(r(k);\delta) \tag{32}$$

where $v=I/K$ and $v_r>0$. The reduced wealth equation is derived by dividing (18) by PK and substituting $i_s^*=i_s+p^e$,

$$w=w^d(f(k),i_s+p^e,\quad 1+w(r-\delta,i_s+p^e)) \tag{33}$$

where $w_1^d>0$, $w_2^d>0$, $w_3^d>0$.

Only the variable i_s prevents (33) from being in reduced form. By deriving the partial inverse function of (33) for i_s in terms of k, w, p^e, we will have the basis for presenting the reduced model. From (33), we get:

$$i_s=i_s(k,w,p^e) \tag{34}$$

Consider i_{s1}. As w and p^e are fixed,

$$0=w_1^df'(k)+w_2^d\frac{\partial i_s}{\partial k}+w_3^d\frac{\partial r}{\partial k}$$

and

$$i_{s1}=\frac{\partial i_s}{\partial k}=-\frac{1}{w_2^d}(w_1^df'(k))<0$$

Similarly,

$$i_{s2}=\frac{1-w_3^d}{w_2^d}$$

$$i_{s3}=-1$$

Assuming further that $0<w_3^d<1$ implies that i_{s2} is positive.

We have thus completed the prerequisites and can now present the reduced differential dynamic model. Assuming equilibrium in the labor market by (4), (5), and (22), we get:

$$\frac{\dot{k}}{k}=n+\alpha-v-\gamma(s-v) \tag{35}$$

and

$$\dot{w}=\epsilon-[p+v+\gamma(s-v)]w \tag{36}$$

From (35), (32), and (31), we find:

$$\begin{aligned}\dot{k}=[n+\alpha-v(r(k);\delta)-\gamma(s(f(k),\\ i_s(k,w,p^e),1+sq(r(k)-\delta,\\ i_s(k,w,p^e)+p^e))-v(r(k);\delta))]k\end{aligned} \tag{35a}$$

Similarly, from (36) and (27e) we get:

$$\begin{aligned}\dot{w}=\Phi(r,i_s)+v+\phi(s-v)\\ -[\lambda_p(v-s)+v-\gamma(s-v)]w\end{aligned} \tag{36a}$$

Rearranging terms and expressing explicit functional forms gives:

$$\begin{aligned}\dot{w}=\Phi(r(k),i_s(k,w,p^e)+v(r(k);\delta)\\ \cdot[1-\phi-(1+\gamma)w]+(\phi+\gamma w)\\ \cdot[s(f(k),i_s(k,w,p^e),1+sq(r(k)-\delta,\\ i_s(k,w,p^e)+p^e))]-\lambda_p(v-s)w\end{aligned} \tag{36b}$$

Finally from (28) and (29), we get:

$$\begin{aligned}\dot{p}^e=\beta(\lambda p(v(r(k);\delta)-s(f(k),i_s(k,w,p^e),\\ 1+sq(r-\delta,i_s(k,w,p^e)+p^e))-p^e)\end{aligned} \tag{37}$$

Equations (35a), (36a), and (37) are three autonomous differential equations in three variables (k, w, p^e). To determine the properties of the stationary solution, we set the differentials equal to zero and solve. This leads to the following results:

(38)[10]
$$n+\alpha=v(v(k);\delta)$$

$$\begin{aligned}\frac{\epsilon}{w}=\lambda p(v(r(k);\delta)-s(f(k),i_s(k,w,p^e),\\ 1+sq(r-\delta),(i_s+p^e)))+v(r(k);\delta)\end{aligned} \tag{39}$$

$$\begin{aligned}p^e=\lambda p(v(r(k);\delta)-s(f(k),i_s(k,w,p^e),\\ 1+sq((r-\delta),(i_s+p^e)))\end{aligned} \tag{40}$$

Thus we find that:

---

[9] The notation s_j for the partial derivative of s with respect to the jth variable will be used throughout the remaining part of the paper.

[10] Since $s=v$ in long-run equilibrium, the later part of the actual capital formation function drops out.

$$(41)\qquad p = p^e = \frac{\epsilon}{w} - n - \alpha = u - n - \alpha$$

since $\epsilon/w = u$ on the growth path.[11] Further, given the equilibrium values of (k, w, p^e), we can find both the real rental and return on securities. Finally, we can make the following statement about the long-run growth path. Real income, capital, securities, and wealth grow at the rate $n+\alpha$, that of the effective labor rate. The rate of inflation and expected inflation is equal to the rate of money creation less that of labor growth. And finally, the real and money wage rates grow at the rate α and $u-n$.

III. Stability[12]

In this section we will investigate the global asymptotic stability properties of the model.[13] A system of differential equations is said to be globally asymptotically stable if it has the property that when a solution for small $t \geqq 0$ exists, then the solution exists for all $t \geqq 0$ and the solution converges to the equilibrium solution as time (t) grows indefinitely large. The criteria to be employed for investigating this property reduce to the following inequality:

$$(42)\qquad f(y)'J(y)f(y) \leqq 0$$

where

$$(42a)\qquad f(y) = (\dot{k}, \dot{w}, \dot{p}^e)$$

the system (35a), (36a), and (38) while

$$(42b)\qquad J(y) = \begin{bmatrix} \dot{k}_1 & \dot{k}_2 & \dot{k}_3 \\ \dot{w}_1 & \dot{w}_2 & \dot{w}_3 \\ \dot{p}_1^e & \dot{p}_2^e & \dot{p}_3^e \end{bmatrix}$$

is its system of partial derivatives. This condition, in other words, is that the quadratic form derived by taking the system of differential equations and its partial derivatives is negative definite or semidefinite.[14] A necessary and sufficient condition for (42) to be negative definite is that the naturally ordered principal minors of $J(y)$ alternate in sign,[15] i.e.,

$$(43)\qquad \dot{k} < 0, \qquad \begin{vmatrix} \dot{k}_1 & \dot{k}_2 \\ \dot{w}_1 & \dot{w}_2 \end{vmatrix} > 0, \qquad |J| < 0$$

To determine whether condition (43) is satisfied, we must first interpret the partial derivatives of system. These partial derivatives follow:

$$(43a)\qquad \dot{k}_1 = k[-v_r r_k - \gamma(s_1 f'(k) + s_2 I_{s1} + s_3(sq_1 r_k + sq_2 i_{s1}) - v_r r_k)] + [n + \alpha - v - \gamma(s - v)]$$

$$(43b)\qquad \dot{k}_2 = k[-\gamma(s_2 i_{s2} + s_3 s q_2 i_{s2})]$$

$$(43c)\qquad \dot{k}_3 = k[-\gamma(s_2 i_{s3} + s_3 s q_2(i_{s3} + 1))]$$

$$(43d)\qquad \dot{w}_1 = \Phi_1(r_k + \Phi_2 i_{s1} + v_r r_k(1 - \phi - (1 + \gamma)w - \lambda_p')w) + (\phi + \gamma w + \lambda_p' w)[s_1 f'(k) + s_2 i_{s1} + s_3(sq_1 r_k + sq_2 i_{s1})]$$

$$(43e)\qquad \dot{w}_2 = \Phi_2 i_{s2} - v + \gamma(s - v) + (\phi + \gamma w + \lambda_p' w)(s_2 i_{s2} + s_3 s q_2 i_{s2}) - \lambda_p(v - s)$$

$$(43f)\qquad \dot{w}_3 = \Phi_2 i_{s3} + (\phi + \gamma w + w\lambda_p') \cdot (s_2 i_{s3} + s_3 s q_2(i_{s3} + 1))$$

---

[11] ϵ/w is the rate of security issues. If that rate was different from u in equilibrium, the portfolio composition would be changing over time, i.e., the share of the control agent would increase or decrease as $\epsilon/u \lesseqgtr u$.

[12] Comparative static properties of the model were not presented. The results of this analysis could be summarized as follows. The effect of increasing the rate of money issuance is to lower capital intensity, a result consistent with Tobin's model. Thus monetary policy in this type of model is nonneutral.

[13] All previous authors in the money and growth area have restricted their analysis to local stability, i.e., investigating the stability of the linear approximations of Taylor series expansions.

[14] See Hartman, pp. 537–40 and particularly note exercise (B) on p. 538 and corollary 11.3 on p. 540. Although this negative definite condition is not sufficient by itself, it will be assumed that the remaining conditions such as uniqueness of solution path given initial conditions are met.

[15] See Hadley, p. 262.

(43g) $\dot{p}_1^{s} = \beta\lambda_p('v_r r_k - s_1 f'(k) - s_2 i_{s1} - s_3(sq_1 r_k + sq_2 i_{s1}))$

(43h) $\dot{p}_2^{e} = \beta\lambda_p('-s_2 i_{s2} - s_3 sq_2 i_{s2})$

(43i) $\dot{p}_3^{e} = \beta(\lambda_p'(-s_2 i_{s3} - ss_3 sq_2(i_{s3}+1)) - 1)$

Given the conditions which have so far been imposed on the system, none of the partial derivatives are of determinate sign. However, with the addition of two plausible conditions all but (43a) become determinate. These two conditions are:

(44a) $-s_2 > s_3 sq_2$

(44b) $0 < -i_{s3} < 1$

To determine the sign of $\dot{k}_1$ additional conditions need to be imposed. Thus,

(44c) $\dot{k}_1 < 0$ if

$$k\gamma[-v_r r_k + s_1 f'(k) + s_3 sq_1 r_k + s_2 i_{s1}] + v - \gamma(s - v) > -k(v_r r_k - \gamma s_3 sq_2 i_{s1}) + n + \alpha$$

This is plausible since for all k greater than k equilibrium, v is greater than $n+\alpha$. It has already been assumed that s_2 is greater than $-s_3 sq_2$ which implies $\gamma s_2 i_{s1}$ is greater than $-\gamma s_3 sq_2 i_{s1}$. Thus the crucial problem is whether the three remaining negative terms are greater than $-v_r r_k$. Although this seems logical, there is no way to argue this on a priori grounds. We find that:

(42c) $$J(y) = \begin{bmatrix} - & + & - \\ + & - & + \\ - & + & - \end{bmatrix}$$

Notice that the symmetry condition is satisfied and that the diagonal elements are negative. Thus the first of the three conditions of (43) is met, i.e., $\dot{k}_1 < 0$.

The remaining two conditions of (43) are guaranteed if the following hold:

(44d) $1 > \gamma(1 + kX_1)$

(44e) $-v_r r_k - \Phi_2 i_{s2} > \Phi_1 r_k$

(44f) $-v_r r_k + \Phi_1 r_k + 2\lambda_p' X_3 > -v_r r_k w + \Phi_2 i_{s2}$

(44g) $\gamma > \lambda_p'$

(44h)[16] $\phi + 2\lambda_p' > 1$

where

$$X_1 = s_1 f'(k) + s_2 i_{s1} + s_3(sq_1 r_k + sq_2 i_{s1})$$

$$X_3 = s_2 i_{s3} + s_3 sq_2(1 + i_{s3})$$

The condition expressed as (44d) implies that the cumulative effect of the forced saving rate times the effect of a change in capital intensity on savings must be less than one. (44e) and (44f) together imply that the effect of changing capital intensity on investment must be greater than the indirect effects through the flow portfolio decisions. Finally, (44g) and (44h) relate to relative adjustment coefficients. Thus the rate of forced saving must be greater than the rate of adjustment in inflation while the total effect of the security market flow adjustment must be greater than one.

Conditions (44a)–(44h) are sufficient to guarantee the stability of the model. These additional conditions are of two kinds. The first set relate to the slopes of saving investment terms and to the relative size of direct as compared with indirect effects. It has been assumed throughout that the direct effects are greater than the indirect effects. The second conditions relate to the relative efficiency of the different markets in the adjustment process. It is a consistent attribute of money and growth models that the adjustment in the rate of inflation tends to be destabilizing. As (44g) implies, it is crucial that the rate of forced saving and by implication the rate of security market adjustment be greater than that of the rate of inflation. Thus a policy which would tend to inhibit changes in the rate of interest would tend to lead to prolonged periods of inflation. Although this is only posed here as a hypothesis, it is one that is well worth further investigation.

---

[16] In addition, there are two more conditions which guarantee that $n+\alpha-v-\gamma(s-v)$ will not dominate the determinant sign. Since these are rather complex terms whose economic meaning is relatively difficult to interpret, they will be omitted from discussion here.

IV. Conclusions

This paper investigated the effects of introducing an equity security into a money and growth framework. The model was developed explicitly in terms of securities and capital instead of the more traditional treatment in terms of money and capital. The model was reduced to three differential equations in capital intensity, the wealth ratio, and price expectations. Based on this reduced system, equilibrium and global stability were analyzed. The stability of the model was found to depend crucially on whether the direct interest rate effect was greater than the offsetting wealth effect and whether the rate of forced saving was greater than the adjustment in the rate of inflation. It is proposed that the crucial role of financial markets in general, and security markets in particular, is to take the short-run impact of economic disturbances and allow the rate of inflation to adjust slowly to changes in the economic environment.

REFERENCES

E. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *J. Finance,* May 1970, *25,* 383–423.

S. Fischer, "Keynes-Wicksell and Neoclassical Models of Money and Growth," *Amer. Econ. Rev.,* Dec. 1972, *62,* 880–90.

H. Grossman, "Money, Interest and Prices in Market Disequilibrium," *J. Polit. Econ.,* Sept./Oct. 1971, *79,* 943–61.

——— **and R. Barro**, "A General Disequilibrium Model of Income and Employment," *Amer. Econ. Rev.,* Mar. 1971, *61,* 82–93.

M. Hadjimichalakis, "Money Expectations and Dynamics, An Alternative View," *Int. Econ. Rev.,* Oct. 1971, *12,* 381–402.

G. Hadley, *Linear Algebra,* Reading 1961.

P. Hartman, *Ordinary Differential Equations,* New York 1964.

G. Horwich, "Open Market Operations, The Rate of Interest, and the Price Level," *J. Finance,* 1955, *10,* 838–39.

———, *Money, Capital and Prices,* Homewood 1964.

D. Jaffee, *Credit Rationing and the Commercial Loan Market,* New York 1971.

D. Levhari and D. Patinkin, "The Role of Money in a Simple Growth Model," *Amer. Econ. Rev.,* Sept. 1968, *58,* 713–53.

K. Nagatani, "A Note on Professor Tobin's 'Money and Economic Growth'," *Econometrica,* Jan. 1970, *38,* 171–75.

D. Purvis, "Short-Run Dynamics in Models of Money and Economic Growth," *Amer. Econ. Rev.,* Mar. 1973, *63,* 12–23.

M. Shane, "Capital Markets and the Process of Economic Growth," unpublished doctoral dissertation, Purdue Univ. 1969; abstracted in *J. Finance,* Sept. 1970, *25,* 950–51.

M. Sidrauski and D. Foley, "Portfolio Choice, Investment, and Growth," *Amer. Econ. Rev.,* Mar. 1970, *60,* 44–63.

J. Stein, " 'Neoclassical' and 'Keynes-Wicksell' Monetary Growth Models," *J. Money, Credit, Banking,* May 1969, *1,* 153–71.

F. Takahashi, "Money Supply and Economic Growth," *Econometrica,* Mar. 1971, *39,* 285–304.

S. Taylor, "Flow of Funds Account, 1945–1968," Board of Governors of the Federal Reserve System, Washington 1970.

J. Tobin, "Money and Economic Growth," *Econometrica,* Oct. 1965, *33,* 671–84.

Labor Supply Under Uncertainty and the Rate of Unemployment

By MICHAEL J. HARTLEY AND NAGESH S. REVANKAR*

In the labor market, the response of supply side phenomena—participation rates, migration decisions and labor supply—to changes in the rate of unemployment has been one of the most heavily investigated research topics in the recent literature.

With respect to participation rates, application of the well-known principles of consumer demand theory to the problem of allocation of time between market work and leisure has suggested the presence of two counterbalancing effects. On the one hand, additional workers may enter the labor force to bolster declining family incomes during a recession. Alternatively, the presence of high unemployment rates may lead unemployed workers to withdraw from the labor force altogether, since search costs may outweigh the potential benefits of employment. In order to ascertain which of these forces is dominant, a vast body of empirical research has arisen.[1] These studies overwhelmingly support the conclusions that the unemployment rate is a significant variable in explaining labor force participation, and that its net effect is negative—indicating that the "discouraged" workers dominate the "additional" workers.

In two recent papers on rural-urban labor migration by Michael Todaro (one coauthored by John Harris), the simple wage differential approach is extended by recognizing the fact that a large pool of unemployed and underemployed urban workers must certainly affect the prospective migrant's probability of finding employment in the urban sector.

In this paper we shall treat the labor supply decision. Though very different in formulation, we propose a model of the individual's labor supply decision which also treats uncertainty as due to the existence of unemployment. We therefore offer a simple *prototype* model to demonstrate why the unemployment rate belongs in the labor supply function. We feel that in providing a rigorous micro-economic foundation for this phenomenon, our paper serves as a useful background piece for much empirical work on labor supply.

I. The Stochastic Foundations of the Model

The neoclassical formulation of the individual's labor supply function is derived from a simple model of choice between income (or consumption) and leisure, wherein a utility function of these variables is maximized subject to linear budget and time constraints.[2] We characterize such models as *certainty* models, firstly because there are no stochastic aspects in their formulation, and secondly because they rely heavily on the competitive market assumption, whereby wages are perfectly flexible and adjust to eliminate any excess supply in the labor market. Here, unemployment is eliminated by the market clearing process, and all offers to work at the "going" wage rate, w, are realized in actual employment; i.e., an individual's wage *income*, wN_i, is immediately calculable (in his budget constraint) once his labor *offer*, N_i, is determined. Clearly once

---

* Assistant and associate professors, State University of New York, Buffalo, respectively. We wish to acknowledge helpful discussions with Murray Brown, Arthur Butler, and Kevin Sontheimer. Hartley also thanks Brown for support under a National Science Foundation grant (for construction of a regional-national econometric model of Italy) under which this contribution arose.

[1] An excellent survey of this literature is given by Jacob Mincer (1966) which summarizes his own work (1962), as well as that of, among others, W. L. Hansen, S. Altman, G. G. Cain, W. G. Bowen and T. A. Finnegan (1965), T. Dernburg and K. Strand, A. Tella, and S. Cooper and D. F. Johnston. See Bowen and Finnegan for most of these references.

[2] See, e.g., James Henderson and Richard Quandt, pp. 23–24 Insofar as such models are standard, we shall not elaborate here.

the competitive assumption is relaxed, and unemployment is introduced as an admissible state, the individual must reckon with the possibility that his wage income may be zero irrespective of his own labor offer. We therefore propose to develop a model along customary lines which incorporates this type of uncertainty.

It is our contention that the unemployment rate is a legitimate explanatory variable in the *structural* labor supply equation. It is, of course, logically possible that a complete model of the labor market—a labor supply, a labor demand, and a wage-adjustment equation—when solved will exhibit the observed relationship between labor supplied and unemployment, without the latter being present in the supply equation itself. Instead, we shall develop a model to justify inclusion of the unemployment rate in the structural hypothesis as a consequence of the uncertainty surrounding the individual's employment status.

Let i be an index for a particular individual who is part of the labor force. We shall characterize the individual's employment status by means of the binary random variable x_i. Thus individual i is either

(i) Unemployed $(x_i = 0)$

or

(ii) Employed $(x_i = 1)$

By definition, every individual in the labor force must be willing to offer work. Let Un_i, $0 < Un_i < 1$, denote the *probability* that individual i will find himself unemployed in the forthcoming planning period.[3] Therefore, x_i has a probability density function given by

$$Pr(x_i = 0) = Un_i \qquad (1)$$
$$Pr(x_i = 1) = 1 - Un_i$$

Let w_i denote the wage rate and u_i denote the unemployment compensation for the ith individual.[4] Define the random variable

$$c_i = c_i(w_i, u_i, x_i, N_i) \qquad (2)$$
$$= (1 - x_i)u_i + x_i w_i N_i$$

as the *total compensation* accruing to individual i. Note that when $x_i = 0$, $c_i = u_i$, whereas if $x_i = 1$, $c_i = w_i N_i$.

In the next section we consider a model to determine the optimal amount of labor services, N_i, to offer. Thus, conditional upon N_i, the *expected* total compensation is given by

$$\mu_{c_i} \equiv E(c_i \mid N_i) \qquad (3)$$
$$= Un_i u_i + (1 - Un_i) w_i N_i$$

Insofar as $w_i N_i > u_i$ it is clear that, given N_i, the expected total compensation, μ_{ci}, varies directly with the anticipated wage rate w_i and inversely with the unemployment rate Un_i.

II. A Model of Labor Supply Under Uncertainty

In the previous section we introduced certain variables which an individual will consider in making a decision about his optimal labor supply. We now turn to the model itself.

A. *The Basic Model*

We shall make the customary assumption that utility is derived from real consumption and leisure. Thus

$$U_i = U_i(C_i, L_i) \qquad (4)$$

where C_i is real consumption and L_i is leisure (nonwork) time of individual i.[5] We also make the usual assumption:

S1: U_i is an ordinal, twice differentiable convex function with continuous first and second partials, with U_{i_1} and $U_{i_2} > 0$.

Note that both the budget and time constraints (which follow) are stochastic, so that C_i and L_i are random variables.

Accordingly U_i of (4) is a random variable, and we shall assume:

---

[3] A particular individual's unemployment rate will typically vary, depending, *inter alia*, on his age, sex, race, marital status, education, etc.

[4] For simplicity we shall treat w_i and u_i as nonstochastic.

[5] The model can easily be generalized to include "future" consumption and leisure. See Robert Lucas and Leonard Rapping.

S2: Each individual maximizes his *expected* utility with respect to his "decision variable" N_i.

Finally, it will be convenient to assume:

S3: The utility function U_i has no certainty bias.[6]

We shall assume that an individual's income arises from three sources. He may earn wage income w_iN_i provided he is employed, and receive unemployment compensation u_i if unemployed. Finally we shall define Y_i as nonwage income from all other sources irrespective of his employment status. Expenditures may be made on consumer goods C_i at a price p.[7] The *budget constraint* is therefore

$$(5)\quad pC_i = c_i + Y_i$$
$$= (1 - x_i)u_i + x_iw_iN_i + Y_i$$

Hence the individual may face two different budget constraints. Either he finds himself unemployed, so that

$$(6)\quad pC_i = u_i + Y_i \quad \text{(with probability } Un_i)$$

or, if his offer to work an amount N_i is realized, his budget constraint is

$$(7)\quad pC_i = w_iN_i + Y_i$$
$$\text{(with probability } 1 - Un_i)$$

Similarly the individual must allocate his time between work and leisure. Let T_i be the total time available.[8] Then the time constraint may be written as

$$(8)\quad T_i = L_i + x_iN_i$$

so that with probability Un_i we have $T_i = L_i$, whereas with probability $1 - Un_i$, $T_i = L_i + N_i$.

The problem for each individual is, therefore, to maximize the conditional expectation of U_i in (4) given N_i with respect to N_i, subject to the random constraints (5) and (8).[9]

B. *The Certainty Equivalent of the Basic Model*

Under assumption S3 it is clear that a solution to the above is also a solution to the problem:

$$(9)\quad \operatorname*{Max}_{N_i} U_i(\overline{C}_i, \overline{L}_i)$$

where the conditional expectations, $\overline{C}_i = E[C_i | N_i]$ and $\overline{L}_i = E[L_i | N_i]$, are both functions of N_i, subject to the constraints,

$$(10)\quad p\overline{C}_i = Un_iu_i + (1 - Un_i)w_iN_i + Y_i$$

and

$$(11)\quad T_i = \overline{L}_i + (1 - Un_i)N_i$$

Equations (10) and (11) arise from taking conditional expectations, given N_i, of (5) and (8), respectively.

Substitution of (11) into (9) leads to the Lagrangean

$$(12)\quad V_i = U_i(\overline{C}_i, T_i - (1 - Un_i)N_i)$$
$$+ \lambda[p\overline{C}_i - Un_iu_i$$
$$- (1 - Un_i)w_iN_i - Y_i]$$

which, upon maximization, leads to the first-order conditions

$$(13)\quad \frac{\partial V_i}{\partial \overline{C}_i} = U_{i_1} + \lambda p = 0$$

$$(14)\quad \frac{\partial V_i}{\partial N_i} = -(1 - Un_i)U_{i_2}$$
$$- \lambda(1 - Un_i)w_i = 0$$

---

[6] A utility function is without certainty bias, see Henri Theil, pp. 420–23, if maximizing $E[U_i(C_i, L_i) | N_i]$ leads to the same value of N_i as maximizing $U_i(E[C_i | N_i], E[L_i | N_i])$. In this case the latter problem (maximizing the utility function of the expected values) is said to be the *certainty equivalent* of the former problem (maximizing the expected value of the utility function). Theil shows that a general form of *quadratic* utility function is without certainty bias.

[7] Saving may be regarded as implicit in the model by regarding it as one of the goods present in the composite commodity C_i having an implicit price of 1 in the consumer price index p. Since we are not concerned here with intertemporal problems, we have ignored this distinction.

[8] T_i may vary with i if instead we interpret i as referring to the household, where there are varying numbers of working-age persons.

[9] This is a stochastic mixed continuous-integer programming problem. Even if the utility function is known it can only be solved numerically. To obtain analytic solutions we must invoke S3.

$$(15)\quad \frac{\partial V_i}{\partial \lambda} = p\bar{C}_i - Un_i u_i - (1 - Un_i)w_i N_i - Y_i = 0$$

We note that solution of (13)–(15) for N_i leads to a labor supply function for individual i,

$$(16)\quad N_i = f_i(p, w_i, u_i, Y_i, Un_i)$$

From (10) it may be verified that N_i *is almost homogeneous of degree zero* in p, u_i, w_i and Y_i.[10] Using C, consumption goods, as the numeraire, we may write (16) as

$$(17)\quad N_i = f_i\left(1, \frac{w_i}{p}, \frac{u_i}{p}, \frac{Y_i}{p}, Un_i\right)$$

We therefore have the following proposition:[11]

PROPOSITION 1: *If the individual maximizes his expected utility, his labor supply depends on real wages, real unemployment compensation, real nonwage income, and his own unemployment rate.*

In order to evaluate the signs of the partials of N_i we proceed in customary fashion by totally differentiating (13)–(15) and using Cramer's Rule to solve the resulting system of three equations in $d\bar{C}_i$, dN_i and $d\lambda$. Let D be the determinant of the resulting bordered Hessian and let D_{ij} be the cofactor of the i, jth element. It may be verified that

$$(18)\quad D = (1 - Un_i)^2[2w_i pU_{i_{12}} - p^2U_{i_{22}} - w_i^2U_{i_{11}}] > 0$$

$$(19)\quad D_{12} = -(1 - Un_i)pw_i < 0$$

$$(20)\quad D_{22} = -p^2 < 0$$

$$(21)\quad D_{32} = (1 - Un_i)(w_iU_{i_{11}} - pU_{i_{12}})$$

and $\lambda < 0$ is the negative of the "marginal utility of money." Equation (18) holds due to the second-order conditions for an interior maximum. Further, it follows that[12]

$$(22)\quad \frac{\partial N_i}{\partial Un_i} = -N_iU_{i_{12}}\frac{D_{12}}{D} + N_i(1 - Un_i)U_{i_{22}}\frac{D_{22}}{D} + (u_i - w_iN_i)\frac{D_{32}}{D}$$

To analyze the signs of these partials it will be convenient to decompose them into the sum of substitution and income effects; i.e., for Un_i,

$$(23)\quad \frac{\partial N_i}{\partial Un_i} = \left(\frac{\partial N_i}{\partial Un_i}\right)_{U_i=\text{constant}} + \Big(\text{Income Effect}\Big)_{\partial N_i/\partial Un_i}$$

To obtain the former we totally differentiate (9) and set $dU_i = 0$. Upon insertion of the first-order conditions (13) and (14), we obtain

$$(24)\quad pd\bar{C}_i + w_i[N_i dUn_i - (1 - Un_i)dN_i] = 0$$

which, in conjunction with the total differential of the budget constraint, yields

$$(25)\quad \bar{C}_i dp - Un_i du_i - (1 - Un_i)N_i dw_i - u_i dUn_i - dY_i = 0$$

---

[10] A function f is *almost homogeneous of degree k* in vector arguments of x and y if it satisfies

$$f(\lambda x, y) = \lambda^k f(x, y)$$

for scalar $\lambda > 0$. See J. Aczel for a discussion of almost homogeneous functions. In our case, let $x = [Un_i]$ and $y = [p, w_i, u_i, Y_i]$. Then clearly f_i of (16) is almost homogeneous of degree zero.

[11] We note that if the individual's anticipated wage is a random variable or if nonwage income is random then the labor supply will depend on expected real wages and expected real nonwage income.

[12] For completeness we also note that

$$\frac{\partial N_i}{\partial Y_i} = \frac{D_{32}}{D}$$

$$\frac{\partial N_i}{\partial u_i} = Un_i\frac{D_{32}}{D}$$

$$\frac{\partial N_i}{\partial p} = -\lambda\frac{D_{12}}{D} - \bar{C}_i\frac{D_{32}}{D}$$

$$\frac{\partial N_i}{\partial w_i} = \lambda(1 - Un_i)\frac{D_{22}}{D} + N_i(1 - Un_i)\frac{D_{32}}{D}$$

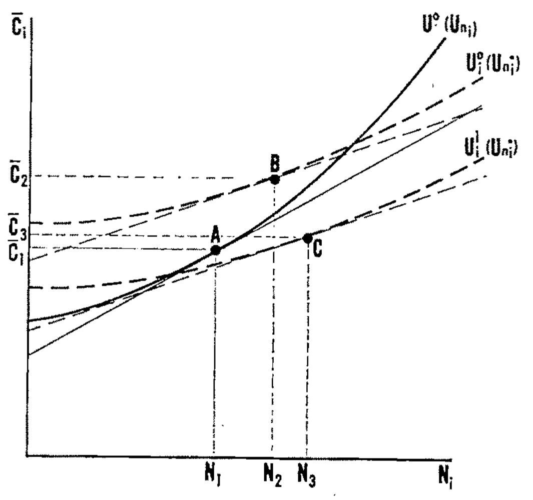


FIGURE 1

It can easily be seen that using (25) the resulting substitution effect in (23) is given by

$$\left(\frac{\partial N_i}{\partial Un_i}\right)_{\overline{U}_i} = \frac{N_i}{1-Un_i} > 0 \tag{26}$$

which leaves as the income effect

$$\left(\text{I.E.}\right)_{\partial N_i/\partial Un_i} = u_i\frac{D_{32}}{D} \tag{27}$$

a term of ambiguous sign.

We should note here that insofar as Un_i is present in the second argument of the utility function, $T_i-(1-Un_i)N_i$, a change in Un_i causes not only a change in the slope and intercept of the budget constraint, but also shifts the entire indifference map. Consequently, the substitution effect, in response to a change in Un_i, captures a movement from an indifference curve in the original map to another in the new map with the same ordinal measure.[13] We illustrate the comparative statics of this situation in Figure 1. Associated with the unemployment rate Un_i we denote the equilibrium by A, which yields utility $U_i^0(Un_i)\equiv U_i(\overline{C}_1,\ T_i-(1-Un_i)N_1)$. If the unemployment rate increases to Un_i^*, the new equilibrium is given by C with utility index $U_i^1(Un_i^*)\equiv U_i(\overline{C}_3,\ T_i-(1-Un_i^*)N_3)$. The movement from A to C is composed of the substitution effect $A\rightarrow B$ plus the income effect $B\rightarrow C$. At point $B=(\overline{C}_2,\ N_2)$ the utility index $U_i^0(Un_i^*)$ in the new indifference map is identical to $U_i^0(Un_i)$ of the original map associated with A. The positive sign of the substitution effect is illustrated since $N_2>N_1$. It indicates that a higher rate of unemployment will induce the individual to offer to work more to preserve the same level of utility.

In the case of the remaining effects, use of (25) and the foregoing analysis indicates that (in fn. 12) $\partial N_i/\partial Y_i$ and $\partial N_i/\partial u_i$ are pure income effects, whereas for $\partial N_i/\partial p$ and $\partial N_i/\partial w_i$ the first terms are substitution effects and the second terms are income effects. In short all substitution effects are of unambiguous sign, and all income effects depend upon the sign of D_{32}. Clearly, if leisure is a normal good, work is inferior and $D_{32}<0$; and conversely for $D_{32}>0$.

We summarize the signs of the partials in Table 1. In the normal case where N_i is inferior (columns (1) and (2)) only the net effects for Y_i and u_i are unambiguous. If the income effect is of either sign but *small*—i.e., dominated by the substitution effect—then the signs of the net effects for p, w_i and Un_i are unambiguous. In general we shall assume that *the income effect is negative but small*. In this case, consistent sign patterns require that:

TABLE 1—SIGNS OF PARTIAL DERIVATIVES OF LABOR SUPPLY

		Net Effect			
Partial of N_i with respect to	Substitution Effect	(1) $D_{32}<0$ and small	(2) $D_{32}<0$ and large	(3) $D_{32}>0$ and small	(4) $D_{32}>0$ and large
Un_i	+	+	−	+	+
w_i	+	+	−	+	+
p	−	−	+	−	−
Y_i	None	−	−	+	+
u_i	None	−	−	+	+

---

[13] This analysis is similar to that of Peter Kalman where prices enter the utility function in consumer demand theory.

PROPOSITION 2: *The individual's labor supply varies directly with real wages and the unemployment rate, and inversely with real unemployment compensation and real nonwage income.*

Finally we note that in the case where the income effect (27) is negligible, i.e., $\partial N_i/\partial Un_i \approx N_i/(1-Un_i)$, we have

$$\frac{\partial^2 N_i}{\partial Un_i^2} \approx \frac{N_i}{(1 - Un_i)^2} > 0 \tag{28}$$

In this case we have the proposition:

PROPOSITION 3: *The higher the level of the unemployment rate Un_i, the greater the rate of response of the individual s labor supply N_i to a change in the unemployment rate.*

III. Summary

In this paper we have attempted to construct a simple prototype model for the micoeconomic foundations of the aggregate labor supply function. We have argued that the standard neoclassical model is inappropriate when there is uncertainty surrounding the individual's decision. A simple characterization of uncertainty in terms of the individual's employment status is sufficient to illustrate the difficulty. This results in a utility maximization problem where the constraints are random. We have shown that, if the utility function is without certainty bias and the individual maximizes expected utility, an analytic solution is possible. We find a labor supply function dependent on (expected) real wages, (expected) real unemployment compensation, (expected) real nonwage income and the unemployment rate. Further, we analyze the signs of the partial derivatives and obtain expressions for the substitution and income effects. We view our results as providing a formal justification for the inclusion of the unemployment rate in the labor supply function.

REFERENCES

J. Aczel, *Lectures on Functional Equations and Their Applications*, New York 1966.

W. G. Bowen and T. A. Finnegan, *The Economics of Labor Force Participation*, Princeton 1969.

J. R. Harris and M. P. Todaro, "Migration, Unemployment, and Development: A Two Sector Analysis," *Amer. Econ. Rev.*, Mar. 1970, *60*, 126–42.

J. M. Henderson and R. E. Quandt, *Microeconomic Theory*, 1st ed., New York 1958.

P. J. Kalman, "Theory of Consumer Behavior When Prices Enter the Utility Function," *Econometrica*, July 1968, *36*, 497–510.

R. E. Lucas and L. A. Rapping, "Real Wages, Employment, and Inflation," *J. Polit. Econ.*, Sept. 1969, *77*, 721–54.

J. Mincer, "Labor Force Participation of Married Women," in *Aspects of Labor Economics*, Universities Nat. Bur. Econ. Res. Conference Series, Princeton 1962, 63–97.

———, "Labor Force Participation and Employment, A Review of Recent Experience," in R. A. Gordon and M. Gordon, eds., *Prosperity and Unemployment*, New York 1966.

M. P. Todaro, "A Model of Labor Migration and Urban Development in Less Developed Countries," *Amer. Econ. Rev.*, Mar. 1969, *59*, 138–48.

Default Risk, Homemade Leverage, and the Modigliani-Miller Theorem

By DAVID P. BARON*

Franco Modigliani and Merton Miller (M-M) have shown in their classic paper that in the absence of taxes and transactions costs the value of a firm does not depend on its debt-equity ratio, since an individual investor may create his own homemade leverage to offset any particular debt-equity ratio chosen by the firm. The M-M theorem recently has been reexamined by Joseph Stiglitz and by Vernon Smith, who consider the implications of less restrictive assumptions than those in the original proof. Stiglitz has indicated that the M-M theorem holds under more general conditions, while Smith focused on preferences of risk averse, expected-utility maximizing investors for various debt-equity ratios for firms, when the probability of default on corporate bonds is positive. In this paper the effects of default risk and differences in the nominal interest rates for firm and individual borrowing are considered for classes of expected-utility maximizing investors who may or may not be risk averse. The paper provides an interpretation of the M-M theorem using the concept of stochastic dominance studied by Josef Hadar and William Russell, Giora Hanoch and Haim Levy, and the authors cited therein.

A distribution function G_1 is said to dominate a distribution function G_2 for the class of increasing utility functions if $G_1(z) \leq G_2(z)$ for all z and $G_1(z^0) < G_2(z^0)$ for some z^0. This dominance criterion implies that for all strictly increasing utility functions U the expected utility with G_1 is greater than that with G_2, or

$$\int_{-\infty}^{\infty} U(t)dG_1(t) > \int_{-\infty}^{\infty} U(t)dG_2(t)$$

Every individual with such a utility function prefers G_1 to G_2, and if all individuals with increasing utility functions prefer G_1 to G_2, then $G_1(z) \leq G_2(z)$ for all z, and $G_1(z^0) < G_2(z^0)$ for some z^0. A distribution G_1 dominates G_2 for the class of nondecreasing, concave utility functions if and only if $\int_{-\infty}^{z}[G_1(t)-G_2(t)]dt \leq 0$ for all z and $G_1 \neq G_2$ for some z_0. Then no risk averse or risk neutral individual prefers G_2 to G_1.

I. The M-M Theorem and Default Risk

To investigate the M-M theorem, consider two firms in the same risk class[1] with identical probability distributions of gross earnings[2] X such that for any state that occurs, both firms have the same earnings. The probability of bankruptcy ($X < 0$) is assumed to be positive (but less than one), and hence the probability of default on any corporate bonds issued is positive. Stochastic constant returns to scale are assumed, and the distribution function of gross earnings is denoted by $F(X)$. It suffices to assume that one firm is financed solely by equity and the other by debt and equity. The value V_1 of the unlevered firm is $V_1 = E_1$ where E_1 is the value of the firm's shares of common stock, while the value V_2 of the levered firm is the sum of the value E_2 of its equity and the value of its outstanding bonds D_2 or $V_2 = E_2 + D_2$. The gross (nominal) interest rate $r (r > 1)$ on bonds already issued is assumed to be a constant. If gross earnings X are less than or equal to the interest plus the repayment of the borrowing, the earnings accrue to bondholders. The firms' investments are assumed to be fixed so that the effects of financing decisions may be isolated. In the absence of transactions costs and taxes, the values of both firms may be shown to be equal, independent of the capital structure of the firm, if investors and the firms may borrow at the same rate r or if all equity owners of the levered firm hold bonds with interest rate r of firms in the same risk class. The only restriction placed on investors' utility functions is that they be increasing in the return from investment.

---

* Associate professor of managerial economics and decision sciences, Northwestern University.

[1] If X_1 and X_2 denote the gross earnings of two firms, respectively, then both are in the same risk class if $X_1 = kX_2$ with probability one. To simplify the notation, assume $k = 1$, and let $X_1 = X_2 = X$.

[2] This representation of the earnings of firms is discussed by Modigliani and Miller and by Jan Mossin, p. 65.

To show that the values of the levered and unlevered firms are equal, first assume that the value V_1 of the unlevered firm exceeds the value V_2 of the levered firm and that an investor holds an α proportion of the equity $V_1=E_1$ of the unlevered firm. The return Y_1 on the investment αV_1 is

$$(1)\qquad Y_1=\begin{cases}0 & \text{if } X\leq 0\\ \alpha X & \text{if } X>0\end{cases}$$

If the investor sells his equity αV_1 in the unlevered firm and invests the proceeds in the equity and bonds of the levered firm in proportions E_2/V_2 and D_2/V_2, respectively, the return Y_2 from the levered firm is[3]

$$(2)\qquad Y_2=\begin{cases}0 & \text{if } X\leq 0\\ \alpha(V_1/V_2)X & \text{if } X>0\end{cases}$$

Let the distribution functions of Y_1 and Y_2 be denoted by G_1 and G_2, respectively. If the firms become bankrupt or have zero earnings, the investor's return on either investment is zero, and hence the probabilities $G_1(0)$ and $G_2(0)$ of no return are equal. With positive earnings and $V_1>V_2$, the probability of obtaining less than some return Y is less with G_2 than with G_1 for all $Y>0$, $[G_2(Y)=F((Y/\alpha)(V_2/V_1))<G_1(Y)=F(Y/\alpha)]$, so the distribution G_2 dominates the distribution of G_1 for the class of increasing utility functions[4] as indicated in Figure 1. If the value of the unlevered firm exceeds that of the levered firm, all investors with increasing utility functions prefer to sell their securities in the unlevered firm and purchase equity and bonds in the levered firm. The value V_2 of the levered firm will be bid up until it is at least as great as the value of the unlevered firm.

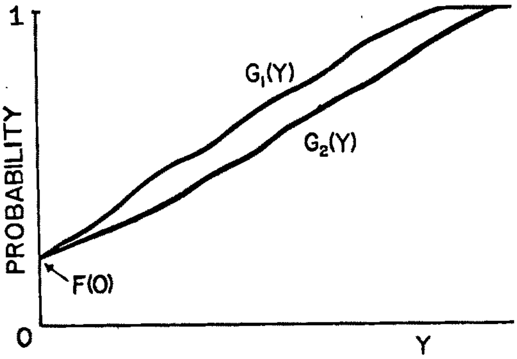


FIGURE 1

The rest of this paper deals with the issue of whether or not the value of the levered firm may exceed the value of the unlevered firm. To demonstrate that the values will be equal when the default risk is positive, assume that $V_2>V_1$ and that all equity investors in the levered firm also hold bonds with a nominal interest rate r of firms in the same risk class as the firms in question. If an investor holds an α fraction of the equity of the levered firm and α^*D_2 bonds, the return on the investment in the levered firm is

$$(3)\qquad Y_2=\begin{cases}0 & \text{if } X\leq 0\\ \alpha^*X & \text{if } 0<X\leq rD_2\\ \alpha X+(\alpha^*-\alpha)rD_2 & \text{if } X>rD_2\end{cases}$$

Let $\alpha^0=\min\{\alpha,\alpha^*\}$, and consider selling $\alpha^0(E_2+D_2)=\alpha^0V_2$ of the investment in the levered firm and buying an (α^0V_2/V_1) portion of the unlevered firm. The return Y_1^0 on the investments in firms two and one then is

$$(4)\qquad Y_1^0=\begin{cases}0 & \text{if } X\leq 0\\ \alpha^*X+\alpha^0X((V_2/V_1)-1) & \\ & \text{if } 0<X\leq rD_2\\ \alpha X+(\alpha^*-\alpha^0)rD_2+\alpha^0X((V_2/V_1)-1) & \\ & \text{if } X>rD_2\end{cases}$$

---

[3] For $X>rD_2$, $Y_2=\alpha V_1(E_2/V_2)(X-rD_2)/E_2+\alpha V_1(D_2/V_2)(rD_2/D_2)=\alpha(V_1/V_2)X$. For $0<X\leq rD_2$ the profit X accrues to the bondholders, and the investor's share is $\alpha V_1(D_2/V_2)$, so $Y_2=\alpha V_1(D_2/V_2)X/D_2=\alpha(V_1/V_2)X$ if $X>0$.

[4] Hadar and Russell (1971, Theorem 4) prove that if $(V_1/V_2)>1$ and $Y_1\geq 0$, the distribution of $Y_2=(V_1/V_2)Y_1$ dominates that for Y_1 for the class of increasing utility functions.

The probabilities $G_1(0)$ and $G_2(0)$ of no return are equal, but if $V_2 > V_1$ and earnings are positive, $G_1(Y) < G_2(Y)$ for $Y > 0$, since investment in the unlevered firm yields a greater share of the earnings than does investment in the bonds and shares of the levered firm. The distribution of Y_1 dominates that for Y_2 for the class of increasing utility functions, so all equity investors holding bonds of firms in the same risk class prefer to sell bonds and shares of the levered firm and to purchase shares in the unlevered firm if $V_2 > V_1$. If all equity investors hold some bonds of firms in the same risk class, then $V_2 \not> V_1$, and the value of the levered and the unlevered firms will be equal independent of their financing.

If all equity investors in the levered firm do not hold bonds, the M-M theorem is valid given a positive probability of bankruptcy if investors may borrow at the same nominal interest rate as the levered firm. Assume that the value V_2 of the levered firm is greater than the value V_1 of the unlevered firm and that an investor owns an α fraction of the equity of the levered firm and no bonds of firms in that risk class. The total return Y_2 is

$$(5) \quad Y_2 = \begin{cases} 0 & \text{if } X \leq rD_2 \equiv X_2^* \\ \alpha(X - rD_2) & \text{if } X > X_2^* \end{cases}$$

If the investor sells his equity αE_2 of firm two, borrows an amount αD_2 at an interest rate r, and invests $\alpha(E_2 + D_2)$ in the equity of the unlevered firm, the investor owns an $\alpha(E_2 + D_2)/E_1 = \alpha V_2/V_1$ fraction of the equity of the unlevered firm.[5] Assuming that the investor's equity value αV_2 in the unlevered firm may be used as the sole collateral[6] for the personal borrowing αD_2, the investor's total return Y_1 from the unlevered firm is

$$(6) \quad Y_1 = \begin{cases} 0 & \text{if } X \leq rD_2(V_1/V_2) = X_1^* \\ \alpha((V_2/V_1)X - rD_2) & \\ & \text{if } X > X_1^* \end{cases}$$

If $V_2/V_1 > 1$, then $X_1^* < X_2^*$, and the probability $G_1(0) = F(X_1^*)$ of no return from the holdings of shares of the unlevered firm is less than the probability $G_2(0) = F(X_2^*)$ of no earnings from holding shares in the levered firm. For $Y > 0$, $G_1(Y) = F((V_1/V_2) \cdot (Y/\alpha + rD_2)) < G_2(Y) = F(Y/\alpha + rD_2)$ if $V_2 > V_1$, since the investor is able to purchase a larger share of the earnings of the unlevered firm. The distribution function G_1 dominates G_2 for the class of increasing utility functions, so all investors prefer to sell shares of the levered firm, to create homemade leverage by borrowing at an interest rate r, and to invest in the shares of the unlevered firm. Consequently, the values of the levered and unlevered firms are equal, and the cost of capital is independent of the debt-equity ratio.

The above arguments demonstrate that the M-M theorem holds if the probability of firm bankruptcy is positive, if all equity investors in the levered firm hold bonds yielding r of firms in the same risk class, or if they may borrow at the same rate as firms and the borrowing is secured solely by the shares of the unlevered firm.[7] This result is in contrast to Smith's Theorem 2 if all investors hold bonds yielding r of firms in the same risk class. The result is equivalent to Stiglitz's extension, p. 788, of the M-M theorem and to Smith's Theorem 3 for the case in which the investor may borrow at the same rate as firms. Risk aversion is not assumed here, however.

Stiglitz's extension is based on the assumption that if an investor buys, on margin, shares in the unlevered firm using the shares as collateral, he can borrow at the same interest rate r as the levered firm. Smith has argued that margin borrowing will be possible only at a higher interest rate than that at which the firm borrows. To show that the investor can borrow at a rate r using the shares in the unlevered firm as collateral, consider a lender[8] with initial wealth W^*

---

[5] The margin requirement necessary to permit sufficient borrowing is (D_2/V_2).

[6] That is, if the investor defaults, he forfeits his shares of the unlevered firm and none of his other assets.

[7] Stiglitz indicates that the M-M theorem holds if instead of personal borrowing the investor may sell short the shares of the levered firm pledging the new shares of the unlevered firm as the sole collateral.

[8] The distinction between lenders and investors is only made for expository purposes. Any lender may also hold shares of the firms.

who invests αD_2 in bonds of the levered firm. His contingent terminal wealth is

$$(7)\qquad W=\begin{cases} W^*-\alpha D_2 & \text{if } X\leq 0\\ W^*-\alpha D_2+\alpha X & \text{if } 0<X\leq rD_2\\ W^*+\alpha D_2(r-1) & \text{if } X>rD_2\end{cases}$$

Next suppose that the lender lends αD_2 to the investor who invests $\alpha(D_2+E_2)$ in the unlevered firm and pledges the shares as collateral. The collateral arrangement is such that if earnings X are insufficient to cover the debt obligation $\alpha D_2 r$ the return $\alpha(V_2/V_1)X$ (for $X>0$) goes to the lender. The lender's contingent wealth W' then is

$$(8)\quad W'=\begin{cases} W^*-\alpha D_2 & \text{if } X\leq 0\\ W^*-\alpha D_2+\alpha(V_2/V_1)X & \\ & \text{if } 0<X\leq rD_2\\ W^*+\alpha D_2(r-1) & \text{if } X>rD_2\end{cases}$$

If $V_2>V_1$, the lender prefers to lend to the investor rather than hold bonds of the levered firm, since the distribution of W' dominates the distribution of W for the class of increasing utility functions. For $V_2=V_1$, W and W' are identical so the investor is able to borrow at the same nominal interest rate as the firm if the above collateral arrangement can be made.[9] Consequently, if $V_2>V_1$ and the margin permitted is at least D_2/V_2, investors will create homemade leverage by borrowing at a rate r and will purchase shares in the unlevered firm rather than hold shares of the levered firm This process will equate the values of the firms.

II. Differences in Firm and Investor Nominal Borrowing Rates

If the investors are not able to borrow the proportion D_2/V_2 of the investment in the unlevered firm because of margin restrictions, limitations on short sales, or market imperfections, the investor seeking to create homemade leverage will be required to secure personal borrowing with collateral or to pay a higher nominal interest rate than r. The nominal interest rate r^* at which an investor may borrow to create homemade leverage depends on the collateral and may differ from the rate r for the firm. Personal borrowing creates a new security, and the implications of this new security for the M-M theorem are investigated in this section.

Suppose that collateral αC, satisfying $0\leq\alpha C\leq\alpha r^* D_2$, is pledged against the amount borrowed αD_2, where the amount αC may be determined by a specific collateral arrangement or by personal bankruptcy laws. If earnings plus the collateral are insufficient to cover the debt obligations ($X\leq(r^*D_2-C)\cdot(V_1/V_2)$), the investor is assumed to pay the lender ($\alpha(V_2/V_1)X+\alpha C$) (with probability one). If $X>(V_1/V_2)(r^*D_2-C)$, the return plus the collateral covers the debt obligations. A lender's contingent wealth W_1' with such a loan is

$$(9)\quad W_1'=\begin{cases} W^*-\alpha D_2+\alpha C & \\ \qquad \text{if } X\leq 0 & \\ W^*-\alpha D_2+\alpha(C+(V_2/V_1)X) & \\ \qquad \text{if } 0<X\leq(V_1/V_2)(r^*D_2-C) & \\ W^*+\alpha D_2(r^*-1) & \\ \qquad \text{if } X>(V_1/V_2)(r^*D_2-C) & \end{cases}$$

If $r^*\geq r$, $V_2\geq V_1$, and $C>0$, $W_1'\geq W$ for all X and $W_1'>W$ for $X\leq(V_1/V_2)(r^*D_2-C)$. The distribution of W_1 dominates the distribution of W, so all lenders with increasing utility functions prefer to lend to the investor rather than hold bonds in the levered firm. Thus, $r^*\leq r$ if $C>0$.

If the investor sells his equity αE_2 in the levered firm, borrows αD_2 at r^* by pledging collateral αC under the above terms, and invests $\alpha(E_2+D_2)$ in shares of the unlevered firm, the return Y_1^* is

$$(10)\quad Y_1^*=\begin{cases} -\alpha C & \text{if } X\leq(V_1/V_2)(r^*D_2-C)\\ \alpha((V_2/V_1)X-r^*D_2) & \\ & \text{if } X>(V_1/V_2)(r^*D_2-C)\end{cases}$$

For any nominal interest rate r^* the distribution of Y_1^* does not dominate, for the class of increasing utility functions, the distribution of the return Y_2 in (5) on equity ownership in the levered firm.[10] Consequently, the

---

[9] The same result obtains if both firms are levered and the investor invests an (E_1/V_1) fraction in shares and a (D_1/V_1) fraction in bonds.

[10] In this case, an investor may wish to diversify between Y_1^* and Y_2, but the conclusions are essentially the same.

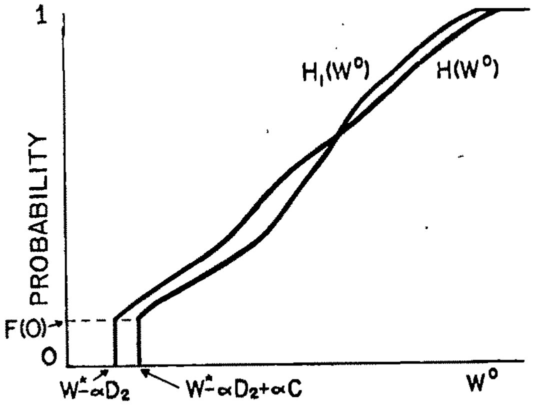


FIGURE 2

value of the levered firm may be greater than the value of the unlevered firm, since all equity owners of the levered firm may not prefer to create homemade leverage and switch to equity ownership in the unlevered firm.

If all lenders have concave utility functions, a further condition on the nominal interest rate r^* may be obtained using the concept of dominance for the class of concave utility functions. For $r^*\le r$ and for $X<(>)X^*\equiv(V_1/V_2)(r^*D_2-C)$ the probability $H(W^0)$ that $W\le W^0$ is greater (less) than the probability $H_1(W^0)$ that $W_1'\le W^0$ for all $W^0<(>)W+\alpha D_2(r^*-1)$[11] as indicated in Figure 2. Given this property, Hanoch and Levy (Theorem 3) have shown that all lenders with concave utility functions prefer the distribution of W_1' to W if and only if $E(W_1')-E(W)\ge 0$, where E denotes expectation. All lenders thus prefer to lend to the investor rather than hold the bonds of the levered firm if, and only if,[12]

$$(11)\quad E(W_1')-E(W)$$

$$=\alpha CF(X^*)+\alpha\int_0^{X^*}X(V_2/V_1-1)dF(X)$$

$$-\alpha\int_{X^*}^{rD_2}XdF(X)$$

$$+\alpha D_2(r^*(1-F(X^*))$$

$$-r(1-F(rD_2)))\ge 0$$

Let r^{**}[13] be defined as the value of r^* such that $E(W_1')-E(W)=0$. If F is absolutely continuous, $\partial[E(W_1')-E(W)]/\partial r^*=\alpha D_2(1-F(X^*))>0$, so for $r^*>r^{**}$ the difference in the expected wealth levels is positive.[14] The interpretation of this result is that the distribution of W_1' is "less risky" than the distribution of W if $r^*\ge r^{**}$, and hence all risk-averse investors prefer H_1 to H. Consequently, if the expected return from lending to investors seeking to create homemade leverage is at least as great as that from holding bonds, all risk averse lenders prefer to make secured loans at a nominal interest rate of r^{**} or higher rather than to hold bonds yielding r. To determine if there is any demand for secured loans at a nominal interest rate r^*, the investor's opportunities must be investigated.

If all investors have concave utility functions, all investors may prefer the distribution of Y_2 in (5) to that of Y_1^* in (10), since the distributions of Y_1^* and Y_2 satisfy the conditions of Hanoch and Levy's Theorem 3. All risk averse investors prefer Y_2 to Y_1^* if and only if $E(Y_2)-E(Y_1^*)\ge 0$. Evaluating the difference in the expected returns yields[15]

$$(12)\quad E(Y_2)-E(Y_1^*)$$

$$=\alpha CF(X^*)-\alpha\int_{X^*}^{rD_2}(V_2/V_1)XdF(X)$$

$$+\alpha\int_{rD_2}^{\infty}X(1-V_2/V_1)dF(X)$$

$$+\alpha D_2(r^*(1-F(X^*))$$

$$-r(1-F(rD_2)))$$

---

[11] This condition implies that the distribution functions of W and W_1' cross only once.

[12] The condition in (11) holds for $X^*\le rD_2$ (equivalently, $r^*\le(V_2/V_1)(r+C/D_2)$). If $X^*>rD_2$, then $r^*>r$ and the distribution of W_1' dominates the distribution of W for the class of increasing utility functions implying that $E(W_1')-E(W)>0$.

[13] The nominal interest rate r^{**} is a function of $(V_2, V_1, C, r, F, D_2, \alpha)$.

[14] The interest rate r^{**} is less than r, since at $V_2=V_1$ and $r^{**}\ge r$, $E(W_1')-E(W)>0$, and $E(W_1')-E(W)$ is increasing in r^* and in V_2 as indicated in (14) below.

[15] The expression in (12) holds for $r^*\le(V_2/V_1)r+C/D_2$. If $r^*>(V_2/V_1)r+C/D_2$, a similar expression may be determined.

where $X^*=(V_1/V_2)(r^*D_2-C)$. If $E(Y_2)-E(Y_1^*)\geq 0$, all investors with concave utility functions prefer to hold their equity in the levered firm rather than to create homemade leverage and switch to equity ownership in the unlevered firm.

To investigate firm valuation when collateral is pledged to create homemade leverage, one must determine if investors holding equity in the levered firm will wish to borrow at r^* and lenders will wish to lend at r^*. All investors and all lenders are assumed to have concave utility functions. The differences in expected earnings opportunities for investors and expected wealth opportunities for lenders are related by

$$E(Y_2) - E(Y_1^*) = E(W_1') - E(W) + \alpha(1 - V_2/V_1)\int_0^\infty XdF(X)$$

If $V_2=V_1$ and $r^*=r^{**}$, $E(Y_2)-E(Y_1^*)=E(W_1')-E(W)=0$. The derivative with respect to r^* of the differences is

$$(13)\quad \left.\frac{\partial[E(Y_2) - E(Y_1^*)]}{\partial r^*}\right|_{V_2=V_1} = \left.\frac{\partial[E(W_1') - E(W)]}{\partial r^*}\right|_{V_2=V_1} = \alpha D_2(1 - F(X^*)) > 0$$

All holders of bonds in the levered firm prefer to lend to equity owners of the levered firm at interest rates $r^*\geq r^{**}$ rather than hold the bonds of the levered firm, but none of those equity owners prefer to borrow, sell their equity in the levered firm, and purchase equity in the unlevered firm. For $r^*<r^{**}$ and $V_2=V_1$, $E(Y_2)-E(Y_1^*)=E(W_1')-E(W)<0$. All equity owners of the levered firm may not prefer to hold their equity, and some may prefer to borrow at $r^*<r^{**}$ to create homemade leverage and switch to ownership in the unlevered firm. For $r^*<r^{**}$ all lenders may not prefer to lend to the equity owners in the levered firm and hence some may prefer to hold bonds in the levered firm.

Alternatively, if $r^*=r^{**}$ with r^{**} defined at $V_2=V_1$, a greater value V_2 increases $E(W_1')-E(W)$ and decreases $E(Y_2)-E(Y_1^*)$, since for all r^*

$$(14)\quad \left.\frac{\partial[E(W_1') - E(W)]}{\partial V_2}\right|_{V_2=V_1} = (\alpha/V_1)\int_0^{X^*} XdF(X) > 0$$

$$\left.\frac{\partial[E(Y_2) - E(Y_1^*)]}{\partial V_2}\right|_{V_2=V_1} = -(\alpha/V_1)\int_{X^*}^{\infty} XdF(X) < 0$$

For $V_2>V_1$ and $r^*=r^{**}$ (defined at $V_2=V_1$), all bondholders with concave utility functions prefer to lend to equity holders in the levered firm, and there may be some equity owners in the levered firm who prefer to borrow to create homemade leverage, sell their equity, and purchase equity in the unlevered firm. The same may occur for certain nominal interest rates $r^*>r^{**}$ if $V_2>V_1$. This process may stop with the value of the levered firm greater than the value of the unlevered firm if investors and lenders may have any concave utility functions. Determination of the equilibrium values requires further assumptions regarding utility functions and/or probability distributions.

As an example, assume that all investors and lenders are risk neutral. If $V_2=V_1$ and $r^*=r^{**}$, no equity holder in the levered firm prefers to sell his equity, borrow, and switch to the unlevered firm, and no bondholder in the levered firm prefers to lend to equity holders in the levered firm. If $V_2=V_1$ and $r^*<(>)r^{**}$, then all (no) risk neutral lenders prefer to hold bonds in the levered firm and all (no) risk neutral equity owners in the levered firm prefer to borrow to create their own leverage. Consequently, if $V_2=V_1$ no transactions will take place, and $V_2=V_1$ and $r^*=r^{**}$ yield an equilibrium. If $V_2>V_1$, then there exists an $r^*>r^{**}$, r^{**} defined for the particular V_2 and V_1 in question, such that $E(W_1')-E(W)>0$ and

$$E(Y_2) - E(Y_1^*) = E(W_1') - E(W) + \alpha(1 - V_2/V_1) \int_0^{\infty} X dF(X) < 0$$

For example, at $r^* = r^{**}$ and $V_2 > V_1$, $E(W_1') - E(W) = 0$ by definition of r^{**} and

$$E(Y_2) - E(Y_1^*) = \alpha(1 - V_2/V_1) \int_0^{\infty} X dF(X) < 0$$

Since $E(W_1') - E(W)$ is continuous in r^*, an $r^* > r^{**}$ may be found such that all bondholders prefer to lend rather than to hold bonds of the levered firm, and all investors prefer to create their own leverage and switch from equity holding in the levered firm to equity holdings in the unlevered firm. This process continues until $V_2 = V_1$ and $r^* = r^{**}$.

If some or all investors or lenders are risk averse, the process may terminate with the value of the levered firm being greater than the value of the unlevered firm. Some risk averse investors may prefer to be protected by the limited liability of the levered firm and will not create homemade leverage by pledging collateral in order to purchase equity in the unlevered firm. Furthermore, some holders of the bonds of the levered firm may not prefer to lend at a low enough interest rate to generate sufficient sales of the shares of the levered firm and purchases of the shares of the unlevered firm to equate the values of the firms.

III. Conclusions

Using stochastic dominance arguments, the M-M theorem has been shown to be valid if all investors are able to borrow at the same nominal interest rate as firms by pledging the securities in the unlevered firm as collateral or if investors in the equity of levered firms also hold bonds yielding r in firms of the same risk class. If because of margin restrictions, for example, the investor is required to pledge additional collateral, the default risk to the lender is reduced, and the nominal interest rate paid by the investor may be less than the nominal interest rate paid by the levered firm. The values of the levered and unlevered firms will be equal if all economic agents are risk neutral, but if investors and lenders are risk averse, the value of the levered firm may be greater than the value of the unlevered firm.

REFERENCES

J. Hadar and W. R. Russell, "Rules for Ordering Uncertain Prospects," *Amer. Econ. Rev.* Mar. 1969, *59*, 25–34.

——— and ———, "Stochastic Dominance and Diversification," *J. Econ. Theor.*, Sept. 1971, *3*, 288–305.

G. Hanoch and H. Levy, "The Efficiency Analysis of Choices Involving Risk," *Rev. Econ. Stud.*, July 1969, *36*, 335–46.

F. Modigliani and M. H. Miller, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261–97.

J. Mossin, *Theory of Financial Markets*, Englewood Cliffs 1973.

V. L. Smith, "Default Risk, Scale, and the Homemade Leverage Theorem," *Amer. Econ. Rev.*, Mar. 1972, *62*, 66–76.

J. E. Stiglitz, "A Re-Examination of the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Dec. 1969, *59*, 784–93.

Patent Life and R&D Rivalry

By MORTON I. KAMIEN AND NANCY L. SCHWARTZ*

Studies of technical advance have focused on incentives to invent, deviations between socially desirable inventive activity and that most profitable for the individual firm, and means of reducing the deviations. These investigations have been piecemeal, centering interest on a single facet or two to the neglect of others. In this paper we will join two nearly independent streams of inquiry to overcome significant shortcomings of each and in so doing, present a single more comprehensive model of firm inventive activity in the face of a patent system.

A patent confers upon its holder an ownership right to a specific product or process for a designated number of years and thereby facilitates collection of a reward for inventive activity. Justification for patent protection rests on the cost of discovery or development being high relative to the cost of imitation. If a discovery can be quickly imitated, the economic gains to the inventor are reduced, dulling the incentive for investment in inventive activity. Patents, by enabling the inventor to appropriate economic benefits of his contribution over a limited period, counteract this tendency toward underinvestment.

Unfortunately, while patent protection acts to correct one source of resource misallocation, it creates another, namely monopoly over the patent's lifetime. Thus an important social variable is the duration of a patent. Recently several investigators have addressed the question of patent life, notably William Nordhaus (1969), Joseph Stiglitz, and F. M. Scherer (1972). Their approach stems from Kenneth Arrow's model for determining incentives to invent. Nordhaus explores the relationship between the size of an invention, measured by the decrease in unit production cost of an existing good, and the life of the patent. He assumes (i) an increasing concave relationship between the research effort required and the cost reduction achieved and (ii) that the inventor selects a level of research activity to maximize the difference between total discounted royalty revenues and the cost of research. The inventor's behavior function relating the magnitude of the invention sought to patent life is then derived. Finally, Nordhaus determines the socially optimal duration of a patent through maximization of a social welfare function, given the inventor's behavior.

Scherer provides a geometric discussion of Nordhaus' theory and suggests additional considerations. Stiglitz' model differs from Nordhaus' in that patent life is determined by maximization of a welfare function subject to the condition that the inventor just break even. Although all this work constitutes a major step towards providing a framework for the evaluation of the patent system, it is incomplete, as the authors recognize.

An essential feature of resource allocation to inventive activity not dealt with in the Arrow-Nordhaus analysis is the consequence of rivalry among potential inventors prior to the granting of a patent. An implicit supposition of their model is that an inventor is certain to be granted a patent. Yet it is difficult to imagine that would-be inventors are not aware that others may be working toward the same technical improvement and that a prior claim to a patent may leave their own efforts unrewarded. Thus in the present study of the relationships among patent life, the speed of development of processes, and the magnitude of invention, an essential role exists for rivalry and concomitant uncertainty about the availability of the patent on completion of the *R&D*.

A second omission in the Arrow-Nordhaus model, related to the first, is recognition of the duration of calendar time in which the research and development takes place. The rapidity with which an improvement is developed has implications for resource allocation. The earlier an innovation is introduced, the sooner do its benefits become available. On the other hand, more rapid development can usually be achieved only at higher cost. Comparison of socially desirable timing with individually profitable timing has been carried out by Scherer (1967) and Yoram Barzel. Incomplete appropriability of profits by the inventor may tend to prolong the development period, while, on the other hand, the benefits of being first tend to accelerate development. In an earlier paper published in 1972, the present authors studied how incomplete appropriability and rivalry enter the firm's determination of the introduction date of an innovation. In that paper, no patent protection was available but imitation was neither immediate nor its timing known in advance. The present paper also extends the earlier work on timing of innovations to include the alternative supposition of a patent system, and to permit the magnitude of invention to be simultaneously variable.

---

* Graduate School of Management, Northwestern University. Financial support of the National Science Foundation is gratefully acknowledged as are the comments of William D. Nordhaus. We retain all responsibility for views expressed.

I. Invention Opportunity and the Cost of *R&D*

It is supposed that product or process improvement not extant or not operational at the initial time zero could yield total net revenues or cost savings of P dollars per unit time when developed. The first firm to complete successful development will receive a patent for its effort. The patent is a right to collect the net revenues of P over the life of the patent, say a period of length L. At the expiration of the patent, rivalry eliminates excess profits. For a firm that discounts continuously at rate r and plans to complete its development at time T, the potential prize has present value

$$\int_{T}^{T+L} e^{-rt} P dt$$

On the other hand, if another firm completes its development first and obtains the patent, then our firm receives no reward for its efforts and loses the development expenditures already incurred.

There may be but one potential invention contemplated, or there may be a range of possible alternative improvements. In either case, we suppose there is no important uncertainty regarding the feasibility of bringing forth the innovation by devoting the requisite amount of effective inventive effort to the *R&D* project. Analysis of the case in which the feasibility of a project is uncertain has been carried out by us in a 1971 paper. In the first instance, for the single invention yielding reward flow $\underline{P}$, the required effective inventive effort is $A = \underline{A}$ so that

$$\text{(1a)} \quad P(A) = \begin{cases} 0 & \text{for } A < \underline{A} \\ \underline{P} & \text{for } A \geq \underline{A} \end{cases}$$

Alternatively, if there is a range of possible inventions, the magnitude or size of invention is a decision variable. The rate of net revenues P from the invention is then assumed to be an increasing concave function of total effective effort:

$$\text{(1b)} \quad P(A) > 0, \quad P'(A) > 0, \quad P'' \leq 0$$

For instance, potential inventions could be means of reducing unit production costs of a good; then a larger invention is one permitting greater production cost savings.

The firm is assumed to be aware of the possibility of rival introduction prior to its own. This anticipation is not focused on a specific rival but rather encompasses the collectivity of potential rivals. The firm's assessment of others' plans is summarized by a subjective probability distribution F, where $F(\mathrm{t})$ denotes the assumed probability that the composite rival will have introduced its improved product or process by time t and claimed the patent.

Since the firm will terminate its development efforts if and when the patent it seeks is claimed by another, the actual pattern of development spending is of consequence. Let $y(\mathrm{t})$ denote the rate of dollar spending for the research project at time t and let $z(\mathrm{t})$ be the cumulative effective effort devoted to the project by time t. The rate of dollar spending and the growth of cumulative effective effort are assumed to be related by a concave monotone function

$$z'(\mathrm{t}) = g(y(\mathrm{t}))$$

where $g \geq 0$, $g' > 0$, $g'' < 0$, for $y \geq 0$ and primes denote derivatives. The concavity of g reflects diminishing returns to the time compression of research, encompassing both classical diminishing returns tendencies and the uncertain and heuristic features of the *R&D* process as discussed in detail by Scherer (1967). Note that uncertainties about details in the *R&D* process are not inconsistent with our supposition of no important uncertainty about the feasibility of successful development within the given time and effort outlines.

Under the foregoing assumptions, the firm chooses a development plan $y(t)$ and completion date T to maximize the present value of its expected profits from the venture

$$(2)\quad \underset{y(t),T}{\text{Maximize}}\,(1 - F(T)) \int_T^{T+L} e^{-rt} P(A)\,dt - \int_0^T e^{-rt} y(t)(1 - F(t))\,dt$$

subject to

$$\int_0^T g(y(t))\,dt = A, \qquad y(t) \geq 0 \quad T > 0$$

Development expenditures are made only until the patent is issued while rewards are collected by the firm only if it is the patent recipient.

Rather than analyze the solution to the problem as stated, we make two additional specializing assumptions that considerably simplify the analysis. Investigation of the specialized problem may be suggestive of patterns to be found with less restrictive assumptions. Specifically, we let

$$(3)\qquad F(t) = 1 - e^{-ht}$$

$$(4)\qquad g(y(t)) = y^a(t), \qquad \text{where } 0 < a < 1$$

The positive parameter $h = F'(t)/(1 - F(t))$ is the instantaneous probability that the composite rival will claim the patent in the next moment of time, given that he has not done so already.

The problem is solved stepwise. We first find the optimal development plan for any fixed magnitude of inventive effort A and introduction date T. Then the optimal A (if variable) and T may be found. Thus we first seek the minimum expected expenditure for a project of size A with planned completion at time T, and with the understanding that spending ceases if and when a rival claims the patent during [0, T]. Inserting assumptions (3) and (4) into (2), we have expected cost

$$(5)\quad C(T, A; h, r) \equiv \min \int_0^T e^{-(r+h)t} y(t)\,dt$$

subject to

$$\int_0^T y^a(t)\,dt = A$$

We can differentiate the Lagrangian form

$$\int_0^T [e^{-(r+h)t} y(t) - \lambda y^a(t)]\,dt,$$

with respect to y, use the constraining relation to solve for the multiplier λ, and finally find that the desired development plan for given T, A is

$$(6)\quad y(t) = [n(r+h)A/(e^{n(r+h)T} - 1)]^{1/a} \cdot \exp\,[(r+h)t/(1-a)], \qquad 0 \leq t \leq T$$

where

$$(7)\qquad n \equiv a/(1-a)$$

The present value of the expected development expenditures is then found from (5) and (6) to be

$$(8)\quad C(T, A; h, r) = A^{1/a}[n(r+h)/(e^{n(r+h)T} - 1)]^{1/n}$$

It can be verified by direct calculation that $C_T < 0$, $C_A > 0$, $C_h < 0$, $C_r < 0$ where subscripts denote partial differentiation with respect to the indicated argument. The present value of expected expenditure on development increases as the development period is contracted and as the amount of effort required increases. It also increases as the discount rate falls, since future expenditures are less heavily discounted. Finally, it rises as the instantaneous conditional probability of rival entry h falls, or equivalently as the expected rival entry date $1/h$ becomes more distant. This follows because the later a rival project is completed, the more development our firm will have been able to do.

II. A Single Potential Invention

We now consider a firm contemplating an innovation the size of which is prespecified; that is, assumption (1a) holds. The firm need only decide whether it wishes to undertake development of the innovation and, if so, to select a schedule for performing the *R&D*. Since the minimum expected cost function (8) is increasing in effort A, it is evident that at most the effort $\underline{A}$ will be employed. Consequently the bars under $\underline{A}$, $\underline{P}$ are occasionally omitted below where the meaning is clear.

The gain to the firm planning development over a period of length T is the net revenue during the life of the patent, if no one else claims the patent before T, and nothing otherwise. Thus the expected present value of the invention to be completed at time T is, from (1a), (2), and (3),

$$e^{-hT}\int_{T}^{T+L} e^{-rt}\underline{P}dt = e^{-(r+h)T}\underline{P}q/r$$

where

$$q \equiv 1 - e^{-rL} \tag{9}$$

The expected cost of the effort has been computed above in (8). The remaining decision for the firm is the planned development period, T. A decision not to undertake the project at all is equivalent to an infinite T. Thus, we seek to

$$\underset{T>0}{\text{maximize}}\; e^{-(r+h)T}Pq/r - A^{1/a}\left[n(r+h)/(e^{n(r+h)T}-1)\right]^{1/n} \tag{10}$$

Setting the derivative of (10) with respect to T equal to zero and solving yields

$$T = -(n(r+h))^{-1} \cdot ln\{1 - A(n(r+h))^{1-a}(r/Pq)^{a}\} \tag{11}$$

The argument of the *log* function in (11) must be positive for T to be well defined; that is, for finite T we need

$$A^{1/a}[n(r+h)]^{1/n} < Pq/r \tag{12}$$

It can be shown that the second derivative of (10), evaluated at (11), is negative so that (11) does maximize (10) providing (12) holds. Whenever the maximum of (10) exists (i.e., (12) holds), it will be positive and equal to

$$\{(Pq/r)^{a} - A[n(r+h)]^{1-a}\}^{1/a}$$

Consequently, (12) is a necessary and sufficient condition for the firm to undertake the project at all. Development will occur only if the reward stream P is large, patent duration L (recall (9)) long, the required *R&D* effort A and discount rate r low. Finally rivalry must not be too intense, or, equivalently, the expected time of rival claim to the patent must be distant. A project can always be discouraged by a short patent life, but even an indefinitely long patent ($q=1$) will not bring forth development effort unless the benefit P is large enough and required effort, interest rate, and rivalry are low enough. Significantly, it is possible that a project be socially worthwhile if undertaken by a monopolist or public agency with no rivalry yet that no single firm would attempt it even with the promise of indefinite patent protection for the successful rival. This conclusion supports the Schumpeterian thesis that monopoly is conducive to technical advance.

The firm will undertake the project if condition (12) holds. The speed of development, or timing of the innovation, will be dependent upon the various parameters, as may be seen by reexamination of (11). In particular, the planned development period will be longer the smaller the value P of the new product or process, the greater the effort A required to obtain the innovation, the shorter patent life L, and the higher the discount rate.

III. A Range of Potential Inventions

Next suppose the firm can choose the magnitude of innovation. Specifically, we suppose that the benefit stream $P(A)$ is an increasing concave function of the amount of effort A; i.e., that (1b) holds. The objective function is the same as in (10) with P replaced by $P(A)$ and A a decision variable. Setting the derivatives of the new (10) with respect to T and A equal to zero gives

$$e^{-(r+h)T}P(A)q/r = A[An(r+h)]^{1/n}e^{n(r+h)T} / [e^{n(r+h)T} - 1]^{1/a} \tag{13}$$

$$(14)\quad e^{-(r+h)\mathrm{T}}P'(A)q/r = [An(r+h)]^{1/n}/a[e^{n(r+h)\mathrm{T}}-1]^{1/n}$$

The second-order conditions for a maximum can be shown to hold whenever (13) and (14) do, so there is at most one solution to these equations. Further, it can be shown that the maximum expected profit from *R&D* is positive so (13), (14) implicitly give the optimal values of invention magnitude A and timing T, if the *R&D* is to be undertaken.

The manner in which the firm's optimizing choice of magnitude and timing depend on the parameters of the problem may be determined from the total differentials of (13) and (14). After some tedious manipulations, it can be shown that the magnitude of invention sought varies directly with patent life and inversely with the discount rate and the intensity of rivalry. The way that timing varies with parametric change is unclear, but depends on the form of the function $P(A)$. In the special case $P(A)=kA^b$, $0<b\leq 1$, the development period is independent of patent life and is longer the lower the interest rate or the degree of rivalry for the patent.

IV. Summary

Heretofore theoretical study of the effect of patent life on innovation has been based on the highly restrictive assumption that there is no rivalry for the patent. We have relaxed that assumption and investigated a firm contemplating an *R&D* project under rivalry. In our study, if the firm is first with the improvement, it gains a patent and associated revenues for a finite period. On the other hand, if a rival completes its *R&D* first and receives the patent, our firm gets nothing, losing whatever investment has been already made. The winner takes all.

We first found the cost minimizing development plan for any project size and development period. Under the specific formulation, the expected expenditure varies directly with the magnitude of improvement and the speed of development, and varies inversely with the degree of rivalry and the discount rate. For the case of a single potential invention of given size, we found a necessary and sufficient condition for the firm to undertake development. In brief, the invention will occur only if the reward stream and duration of its receipt are sufficiently large and the requisite *R&D* effort, discount rate, and rivalry sufficiently small.

When invention size is also subject to choice, it varies directly with patent life and inversely with the discount rate and rivalry. If benefits from the invention can be written as a power function of inventive effort, the firm's choice of development period turns out to be independent of patent life.

Thus, we have sought to extend the Arrow-Nordhaus model of inventive activity and optimal patent life to include the previously neglected but everpresent rivalry among potential innovators, uncertainty, and choice of timing as well as magnitude of inventions. Obviously, further generalizations encompassing additional realities of inventive activity are in order. The plausibility of the conclusions regarding rivalry, timing and magnitude of inventions, and patent life reached with the present model leads us to believe that it provides a solid basis for future investigations.

REFERENCES

K. Arrow, "Economic Welfare and the Allocation of Resources for Invention," in *The Rate and Direction of Inventive Activity*, Princeton 1962, 609–25.

Y. Barzel, "Optimal Timing of Innovations," *Rev. Econ. Statist.*, Aug. 1968, *50*, 348–55.

M. I. Kamien and N. L. Schwartz, "Expenditure Patterns for Risky R and D Projects," *J. Applied Probability*, Mar. 1971, *8*, 60–73.

———, "Timing of Innovations Under Rivalry," *Econometrica*, Jan. 1972, *40*, 43–60.

W. D. Nordhaus, *Invention, Growth, and Welfare*, Cambridge 1969.

———, "The Optimum Life of a Patent: Reply," *Amer. Econ. Rev.*, June 1972, *62*, 428–31.

F. M. Scherer, "Research and Development Allocation Under Rivalry," *Quart. J. Econ.*, Aug. 1967, *81*, 359–94.

———, "Nordhaus' Theory of Optimal Patent Life: A Geometric Reinterpretation," *Amer. Econ. Rev.*, June 1972, *62*, 422–27.

J. Stiglitz, "Discussion," *Amer. Econ. Rev. Proc.*, May 1969, *59*, 46–49.

Life Cycle Consumption and Labor Supply: An Explanation of the Relationship Between Income and Consumption Over the Life Cycle

By JAMES HECKMAN*

In a recent paper in this *Review*, Lester Thurow presents empirical evidence in apparent contradiction with the conventional life cycle consumption theory enunciated by Franco Modigliani and Richard Brumberg, Menahem Yaari, and James Tobin. That theory predicts no necessary relationship between consumption and income receipts at any age, but Thurow demonstrates a strong relationship and shows that income and consumption expenditure both peak in the age interval 45–54. Thurow's principal explanation for his finding is that credit market restrictions prevent consumers from borrowing as much against their future income as they desire at the going interest rate. As long as income tends to increase with age, and discounted future income cannot be fully transferred at the borrowing rate, a consumer's effective net worth increases with age which causes increasing consumption with age. Based on this argument, Thurow recommends government intervention into the consumption loan market to allow for optimal adjustment of consumption.

Keizo Nagatani explains the same facts by building a model based on the uncertainty of future income. By adjusting expected future income for risk, a "typical" consumer will buy less than he would in a riskless environment with the same expected income stream. However, being the typical consumer, he realizes his expected income, and he successively revises his consumption plan upward since his realized income exceeds his risk adjusted income forecast. For this reason, his consumption expenditure and income streams are closely related.

Both authors relax a standard neoclassical assumption to obtain their theoretical results: Thurow assumes imperfect credit markets while Nagatani invokes uncertainty.[1] However, their different explanations lead to different policy implications, since Nagatani's results provide no basis for government intervention to break down institutional barriers in the credit market.[2]

In this paper, we present an alternative neoclassical model which can explain Thurow's results without resort to either credit market imperfections or uncertainty. Rather than treating income as exogenously given, we view earnings as resulting from a life cycle labor supply decision. If individuals are free to set their hours of work, and if wage rates change systematically over the life cycle, the path of consumption of market goods will depend on the wage rate at each age unless goods and leisure are independent of each other in utility.

There is strong empirical evidence that

---

* Columbia University and the National Bureau of Economic Research. This research was sponsored by a U.S. Department of Labor Manpower Administration dissertation grant. I am deeply indebted to Edmund Phelps for comments, and to members of my dissertation committee at Princeton: Orley Ashenfelter, Stanley Black, Richard Quandt, Albert Rees, and Harry Kelejian. I retain responsibility for all errors. This paper is not an official National Bureau publication since the findings reported herein have not yet undergone the full critical review accorded the National Bureau's studies, including approval of the Board of Directors.

[1] Both authors also discuss alternative explanations such as family composition effects, shifts in preferences, and measurement errors.

[2] One might argue that some portion of the risk adjustment of income in the Nagatani model is due to "market imperfection." However, in the presence of uncertainty, market imperfection is not a well-defined operational concept and specific policy recommendations are more difficult to obtain. I am indebted to Phelps for this point.

wage rates vary over the life cycle.[3] We demonstrate below that if the interest rate equals the rate of pure time preference, the level of consumption by age moves exactly with the path of wage rates if time and goods are substitutes in utility. Allowing for a difference between the rate of time preference and the interest rate, an association remains between wage rates and consumption but it is not as precise.

In the first section of this paper, we present an informal statement of the model. The second section is devoted to a rigorous derivation of these propositions.

I

Standard life cycle models assume that the consumer's preferences for goods are the same at each age, and independent of goods consumption at other ages, with future utility discounted at an exponential rate.[4] These models either ignore the consumption of leisure or assume that hours of work are institutionally fixed so that a given lifetime wage path implies an exogenously determined income stream.[5] The only factor creating differences in goods consumption by age is the interest rate net of pure time preference. If the rate of interest exceeds the rate of time preference, a consumer has an incentive to postpone his consumption of goods to later ages.

In our model, we introduce an explicit labor-leisure choice at each age, maintaining the assumption that the utility at one age is independent of the consumption of goods and leisure at other ages. The rates of interest and time preference continue to operate on consumption and leisure in the usual way, but a new factor is added. If the price of leisure is high at certain ages, individuals tend to consume less leisure at those ages.[6]

To focus on this effect, suppose that the interest rate and rate of time preference are zero, and that the price of goods is the same at all ages. If market goods are complements with leisure in the sense that the marginal utility of leisure increases with increments in the consumption of goods,[7] the consumer has an incentive to economize on both his consumption of goods and leisure at ages where the price of leisure is high, since there are gains in utility from consuming time and goods jointly. In this case, he works more and saves more at ages with higher wage rates than at other ages. If market goods are substitutes for leisure in the sense that a reduction in the consumption of leisure raises the marginal utility from consuming goods, at ages where the price of leisure is high relative to other ages, the consumer has an incentive to economize on his leisure but spend more on goods. In this case, at ages where wage rates are high, consumers work more, earn more, and consume more than at ages where wages are lower.

Upon introducing the effect of time preference and interest rates, and assuming that the rate of interest exceeds the rate of time preference, one finds that the consumption of goods and leisure tends to be pushed towards later ages, but the wage-induced pattern of consumption remains, albeit somewhat blurred. We formalize these intuitive statements below.

---

[3] See for example the work of Michael Hurd.

[4] The objections to this utility specification are well known, but it is widely used (see Yaari, Modigliani-Brumberg, Nagatani). For a statement of those objections, see J. Hicks, p. 261.

[5] Frank Ramsey explicitly considers an intertemporal model of consumption and work effort. However, he assumes both intertemporal and contemporaneous additive separability of the preference function in goods and leisure. We demonstrate below that the latter assumption leads to the same predictions as standard models of life cycle consumption which exclude the work decision from consideration.

[6] The same intuitive model is suggested by Milton Friedman, p. 206, and is applied by Robert Lucas and Leonard Rapping in their analysis of the Phillips curve, p. 266.

[7] This definition of complementarity, used by F. Y. Edgeworth, Irving Fisher, and Vilfredo Pareto, differs from the more conventional definition which refers to the sign of substitution effects resulting from the effect of a price change of one good on the consumption of another good. As Paul Samuelson, p. 183, notes, this "direct" definition depends on a cardinal specification of the utility function. Since we follow Modigliani-Brumberg, Yaari, and Tobin in assuming additively separable intertemporal preferences, we have in fact assumed a cardinal specification for the utility function at each age.

II

We assume that the consumer has a strictly concave twice differentiable utility function $U(L(\mathrm{t}), X(\mathrm{t}))$ which is the same at each age (t). $X(\mathrm{t})$ is his consumption of market goods, and $L(\mathrm{t})$ is his consumption of leisure. Two commodities are employed as a simplifying device. Invoking the composite commodity theorem, the argument remains valid if there are many goods whose relative prices do not change with age, and if we introduce the many uses of leisure discussed by Gary Becker. Continuous time is used to facilitate the derivations. All of our results remain valid if time is segmented into discrete periods.

Allowing for time preference at rate ρ, a consumer with horizon T has a lifetime utility function

$$\int_0^{\mathrm{T}} e^{-\rho t} U(L(\mathrm{t}), X(\mathrm{t}))dt \tag{1}$$

At each age, he has a fixed amount of time M available so that if he consumes $L(\mathrm{t})$ units of leisure, his work time is $M-L(\mathrm{t})$. The price of goods at age t is defined to be $P(\mathrm{t})$ while the price of time at age t is $W(\mathrm{t})$, and his money earnings at each age are $W(\mathrm{t})(M-L(\mathrm{t}))$.[8]

Letting r be the rate of interest, $A(0)$ initial assets, and assuming that no constraints are imposed on borrowing or lending except that all loans must be repaid, the consumer's lifetime budget constraint in the absence of bequests is

$$A(0) + \int_0^{\mathrm{T}} e^{-rt}[W(\mathrm{t})(M - L(\mathrm{T})) - P(\mathrm{t})X(\mathrm{t})]dt = 0 \tag{2}$$

The consumer is assumed to maximize (1) subject to (2). Letting U_i be the partial derivative of U with respect to its ith argument, the necessary conditions for an interior maximum are

$$U_1(\mathrm{t}) - \lambda e^{(\rho-r)t} W(\mathrm{t}) = 0 \tag{3}$$

$$U_2(\mathrm{t}) - \lambda e^{(\rho-r)t} P(\mathrm{t}) = 0 \tag{4}$$

and (2), where λ is the Lagrange multiplier associated with constraint (2) and is positive if this constraint is effective. From the strict concavity of $U(\ )$, these conditions are also sufficient for a maximum. (See G. Hadley and Murray Kemp, p. 228.)

One immediate implication of the model is that the discounted lifetime marginal propensity to consume goods out of initial net worth need not be unity. Thus it is not necessary to introduce bequests to achieve this result as Yaari has done.[9] To see this, differentiate equation (2) with respect to $A(0)$ to obtain

$$1 = \int_0^{\mathrm{T}} e^{-rt} P(\mathrm{t}) \frac{\partial X(\mathrm{t})}{\partial A(0)} dt + \int_0^{\mathrm{T}} e^{-rt} W(\mathrm{t}) \frac{\partial L(\mathrm{t})}{\partial A(0)} dt$$

The first term on the right is the discounted lifetime marginal propensity to consume goods out of initial net worth, and it need not be unity as long as $\partial L(\mathrm{t})/\partial A(0) \neq 0$.

It is useful at this point to collect well-known results about strictly concave functions which are needed below. Letting U_{ij} be the second partial derivative of $U(\ )$ with respect to its ith and jth arguments, strict concavity implies

$$\begin{vmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{vmatrix} > 0, \quad U_{11} < 0, \quad U_{22} < 0 \tag{5}$$

while U_{12} is indeterminate in sign. $U_{12} = U_{21}$ from the twice differentiability of $U(\ )$.

From the strict concavity of $U(\ )$, we may solve for $L(\mathrm{t})$ and $X(\mathrm{t})$ as functions of $\lambda e^{(\rho-r)t}P(\mathrm{t})$, and $\lambda e^{(\rho-r)t}W(\mathrm{t})$:

$$L(\mathrm{t}) = L[\lambda e^{(\rho-r)t} W(\mathrm{t}), \lambda e^{(\rho-r)t} P(\mathrm{t})] \tag{6}$$

---

[8] It is possible to introduce human capital accumulation so that more work at one age raises future wages. In this case, the price of time includes the conventional money wage rate and the effect of an extra unit of work effort on discounted future earnings. These considerations complicate, but do not alter, the essential arguments of this paper. A retirement period may also be introduced without altering any of the essential conclusions. For a more complete treatment of these issues, see my dissertation, Essay I.

[9] Ralph Pfouts noted the same point in a one period model.

$$(7)\quad X(t) = X[\lambda e^{(\rho-r)t}W(t),\ \lambda e^{(\rho-r)t}P(t)]$$

Letting L_i and X_i indicate partial derivatives with respect to the ith argument of L and X, respectively,

$$(8)\qquad L_1 < 0, \qquad X_2 < 0$$

while L_2 and X_1 are equal and ambiguous as to sign.

To prove these propositions about the partial derivatives, substitute equations (6) and (7) into equations (3) and (4), and differentiate with respect to the arguments of equations (6) and (7). This manipulation leads to the matrix equation

$$\begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{bmatrix} L_1 & L_2 \\ X_1 & X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Multiplying by the inverse of the first matrix on the left which is guaranteed to exist by the strict concavity of U, we reach

$$\begin{bmatrix} L_1 & L_2 \\ X_1 & X_2 \end{bmatrix} = \frac{1}{D} \begin{bmatrix} U_{22} & -U_{12} \\ -U_{12} & U_{11} \end{bmatrix}$$

where

$$D = \begin{vmatrix} U_{11} & U_{12} \\ U_{12} & U_{22} \end{vmatrix} > 0 \text{ by condition (5).}$$

It is immediately apparent that $L_2 = X_1$, and that if, and only if, $U_{12} = 0$ ($X_1 = L_2 = 0$) will the path of consumption be independent of the wage pattern.

To facilitate the analysis, it is convenient to characterize the wage and price paths by

$$(9)\quad \lambda e^{(\rho-r)t}W(t) = \lambda[W(0) + bm(t)]e^{b(\rho-r)t}$$

$$(10)\quad \lambda e^{(\rho-r)t}P(t) = \lambda P(0)e^{b(\rho-r)t}$$

defining $m(0) = 0$. We assume that the price of market goods ($P(t)$) does not change with age, but we allow for wage growth by introducing the term $m(t)$.

A consumer stationary state is defined as the case where $\rho = r$, and $b = 0$. In this state, it is obvious from equations (6) and (7) that the individual consumes the same leisure ($L(0)$) and market goods ($X(0)$) at all stages of his life cycle.

Suppose we disturb this state by imposing wage growth keeping ρ equal to r. Since we seek to characterize consumer life cycle profiles, it is convenient to normalize all values of leisure and consumption with respect to their initial values $L(0)$ and $X(0)$, respectively.

The effect of this displacement on normalized demand may be written as

$$\left.\frac{\partial\left(\frac{L(t)}{L(0)}\right)}{\partial b}\right|_{b=0} = \frac{1}{L(0)}\left\{L_1(t)\left[W(0)\frac{\partial\lambda}{\partial b} + \lambda m(t)\right] + \left[L_2(t)P(0)\frac{\partial\lambda}{\partial b}\right]\right\} - \frac{L(t)}{L^2(0)}\left[L_1(0)W(0)\frac{\partial\lambda}{\partial b} + L_2(0)P(0)\frac{\partial\lambda}{\partial b}\right]$$

$$\left.\frac{\partial\left(\frac{X(t)}{X(0)}\right)}{\partial b}\right|_{b=0} = \frac{1}{X(0)}\left\{X_1(t)\left[W(0)\frac{\partial\lambda}{\partial b} + \lambda m(t)\right] + \left[X_2(t)P(0)\frac{\partial\lambda}{\partial b}\right]\right\} - \frac{X(t)}{X^2(0)}\left[X_1(0)W(0)\frac{\partial\lambda}{\partial b} + X_2(0)P(0)\frac{\partial\lambda}{\partial b}\right]$$

The partial derivatives are evaluated in the neighborhood of $b = 0$, the stationary state position. Since at the stationary state $L(t) = L(0)$, $L_i(t) = L_i(0)$, $X(t) = X(0)$, and $X_i(t) = X_i(0)$, these expressions may be simplified to

$$\left.\frac{\partial\left(\frac{L(t)}{L(0)}\right)}{\partial b}\right|_{b=0} = \lambda\frac{L_1(t)}{L(0)}m(t)$$

$$\left.\frac{\partial\left(\frac{X(t)}{X(0)}\right)}{\partial b}\right|_{b=0} = \lambda\frac{X_1(t)}{X(0)}m(t)$$

Since $L_1 < 0$, the age with the highest wage rate (i.e., the age with the largest $m(t)$) is the age of minimal consumption of leisure and hence the age of maximal work and earnings.[10]

The sign on X_1 is indeterminate. If goods and leisure are direct complements ($U_{12}>0$), the age for peak wage rates is the age for minimum goods consumption. If time and goods are substitutes ($U_{12}<0$), the age of peak wages and earnings is also the age of peak goods consumption. Only if leisure and goods are independent in utility ($U_{12}=0$) would goods consumption remain the same at all ages when wage rates differ by age, and in this case the predictions of standard life cycle consumption models remain intact.

Simple as this model is, it can account for the observed relationship between earnings and consumption if goods and time are direct substitutes in utility ($U_{12}<0$). Recent empirical work suggests that male hourly wage rates rise to a peak in the age range 45–54 and fall off afterward.[11]

In this simplified model, the only force causing differences in the consumption of leisure and goods by age is the pattern of wage growth. The introduction of a difference between the interest rate and the rate of time preference sets other forces in motion. To fix ideas, suppose wage rates rise *monotonically* with age. If the interest rate exceeds the rate of time preference, the consumer has an incentive to consume more goods and leisure at older ages.[12] But since wage rates increase monotonically with age, the consumer has an incentive to consume less leisure as he ages, and, if goods are complements with leisure, fewer goods with advancing age. In this tug of war, any result can emerge, and in particular it is possible that hours of work and the consumption of goods will reach a peak at an interior age in the life cycle.

To see this more clearly, we again differentiate the normalized demand functions with respect to b in a neighborhood of the stationary state. We now let ρ and r be unequal. After some manipulation we reach

$$(11)\quad \left.\frac{\partial\left(\dfrac{L(\mathrm{t})}{L(0)}\right)}{\partial b}\right|_{b=0} = \frac{\lambda L_1(0)}{L(0)}m(\mathrm{t}) + \frac{\lambda(\rho-r)\mathrm{t}}{L(0)}[L_1W(0)+L_2P(0)]$$

$$(12)\quad \left.\frac{\partial\left(\dfrac{X(\mathrm{t})}{X(0)}\right)}{\partial b}\right|_{b=0} = \frac{\lambda X_1(0)}{X(0)}m(\mathrm{t}) + \frac{\lambda(\rho-r)\mathrm{t}}{X(0)}[X_1W(0)+X_2P(0)]$$

The term in brackets in each expression must be negative if $X(\mathrm{t})$ and $L(\mathrm{t})$ are normal goods.[13] From equation (11), it is seen that

---

[10] In order to cast the theory into observable phenomena, we must replace differentials with differences from the initial stationary state values. Equivalently, we apply the mean value theorem to $L(\mathrm{t})/L(0)$ in a neighborhood of the stationary state path. Thus $(d/db)(L(\mathrm{t})/L(0))db$ becomes $\Delta(L(\mathrm{t})/L(0))$.

[11] Hurd finds a peak for the wage rates of white males in the age range 45–54 using the one in a thousand Census data for 1959, and the Survey of Economic Opportunity (*SEO*) data for 1966. See his Table 3, p. 194. Using the *SEO* data, we regressed hourly wage rates on schooling, age, and age squared, allowing for interactions between schooling and age. We found a peak for hourly wage rates at ages 48–50 for males with 10–12 years of schooling with an approximate standard deviation of four years. Thurow reports a similar age peak for income.

[12] See H. G. Lewis.

[13] To see this, differentiate equations (6) and (7) with respect to $A(0)$ to obtain

$$\frac{\partial L(\mathrm{t})}{\partial A(0)} = e^{(\rho-r)\mathrm{t}}[L_1W(\mathrm{t})+L_2P(\mathrm{t})]\frac{\partial\lambda}{\partial A(0)}$$

$$\frac{\partial X(\mathrm{t})}{\partial A(0)} = e^{(\rho-r)\mathrm{t}}[X_1W(\mathrm{t})+X_2P(\mathrm{t})]\frac{\partial\lambda}{\partial A(0)}$$

Since $\partial\lambda/\partial A(0)<0$, and the assumption of normality implies that the partial derivatives on the left must be positive, the expressions in brackets must be negative. To see why $\partial\lambda/\partial A(0)<0$, we note that

$$\frac{\partial}{\partial A(0)}\left[\int_0^{\mathrm{T}} e^{-\rho\mathrm{t}}U(L(\mathrm{t}),X(\mathrm{t})d\mathrm{t}\right]=\lambda$$

Then

$$\frac{\partial\lambda}{\partial A(0)} = \int_0^{\mathrm{T}} e^{-\rho\mathrm{t}}\begin{bmatrix}\dfrac{\partial L(\mathrm{t})}{\partial A(0)} & \dfrac{\partial X(\mathrm{t})}{\partial A(0)}\end{bmatrix}\begin{bmatrix}U_{11} & U_{12}\\ U_{12} & U_{22}\end{bmatrix}\begin{bmatrix}\dfrac{\partial L(\mathrm{t})}{\partial A(0)}\\[2ex] \dfrac{\partial X(\mathrm{t})}{\partial A(0)}\end{bmatrix}d\mathrm{t}<0$$

since the strict concavity of U implies that the quadratic form inside the integral is negative definite.

the age of peak wage rates is not necessarily the age of peak work effect, unless $\rho = r$.

To pursue the example of monotonic wage growth, suppose that $r > \rho$, and that m(t) increases with age (t). It is possible that the age of peak work effort occurs at a boundary of the life cycle, i.e., 0 or T. But if a peak comes in the interior of this interval, the age of peak work effort comes before the age of peak earnings since earnings are the product of hours worked and wage rates, and the latter increase monotonically with age by assumption.

In this example, only if X_1 is negative ($U_{12} > 0$) and suitably strong will we observe a peak in the consumption of goods at an interior age of the life cycle. If $r < \rho$, we observe an interior peak only if X_1 is positive ($U_{12} < 0$).

If wage rates rise to a peak and decline after a certain age, as Hurd's research suggests, the analysis is only slightly more complicated. Again, it is possible to observe peaks for hours worked and the consumption of goods only at the boundaries of the life cycle. If these peaks occur at interior ages, and $r > \rho$, the peak age for hours worked comes before the age of peak wage rates, and the age of peak earnings occurs between these peak ages.[14]

It is instructive to compare the ordering of the peak ages for hours worked, consumption, earnings, and wage rates, in a simplified model where the interest rate equals the rate of time preference, to a model where they differ. In the first model, the peak age is the same for all variables if $X_1 > 0 (U_{12} < 0)$. Allowing for a difference between the interest rate and the rate of time preference, and assuming $r > \rho$, the peak age for hours of work comes before the peak age for wage rates if $X_1 > 0$.

In fact, it is observed in the Survey of Economic Opportunity (*SEO*) data that the peak age for hours of work occurs in the age interval 39–44 (Heckman, Essay II, p. 24). This comes before the peak age for wage rates and earnings which in these data occurs in the age interval 45–54. Given Thurow's finding on the peak age for the consumption of goods, the data appear to be broadly consistent with a model of $r > \rho$, $X_1 > 0 (U_{12} < 0)$, with a peak in wage rates occurring in the age interval 45–54.

However, this is only one possible explanation of Thurow's facts. We have already shown that even if wage growth is monotonic, it is possible to have a peak in consumption and hours of work in the middle years of the life cycle. More exotic patterns for wage rates, interest rates, and time preferences can easily produce the same results, as can changes in the pattern of goods prices by age.

Without a more extensive empirical analysis of wage patterns it is impossible to be more precise about the exact set of assumptions about preferences necessary to explain Thurow's results. Nevertheless, we can unequivocally assert that Thurow's findings are consistent with a model of perfect certainty and perfect credit markets if consumers face anticipated changing wage rates over their life cycle, and are free to choose their hours of work.

III. Conclusions

In this paper, we have shown that Thurow's empirical results are consistent with a neoclassical model of life cycle consumption and labor supply. It is not necessary to resort to credit market restrictions or uncertainty to explain his facts.

With this in mind, we join Nagatani in questioning Thurow's conclusions in favor of government intervention in the loan market to ensure optimal consumption patterns. Since labor supply behavior and uncertainty can also explain Thurow's empirical results, we must sort out the relative importance of each of these factors in determining observed

---

[14] To prove this, assume that wage rates, W(t), and hours worked, h(t), are continuous functions. From equation (11) it is clear that if $r > \rho$ the peak age for hours of work cannot occur after the peak age for wage rates. The age of peak wage rates (t_3) is implicitly defined by $W'(t_3) = 0$, where X' denotes the derivative with respect to t. The age of peak hours worked (t_1) is defined by $h'(t_1) = 0$. The age of peak earnings (t_2) is defined by

$$[W(t)h(t)]' = 0 = \frac{W'(t)}{W(t)} + \frac{h'(t)}{h(t)}$$

Then clearly $t_1 < t_2 < t_3$.

consumption patterns before firm policy prescriptions are possible.

Throughout this paper, we assume exogenous wage growth, and a single earner. The same predictions developed in this paper emerge from a more general model with wage growth due to human capital investment. In such a model, the "shadow price" of time plays the role of the market wage in this paper, and differs from the observed market wage because work at one age may affect future earnings. (See Heckman, Essay I.) It is also relatively straightforward to generalize our results to multiple worker households, although few new analytical insights emerge. The resulting life cycle consumption path depends on the wage paths of all earners in the household.

REFERENCES

G. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493–517.

M. Friedman, *Price Theory, A Provisional Text*, Chicago 1962.

G. Hadley and M. C. Kemp, *Variational Methods in Economics*, Amsterdam 1971.

J. Heckman, "Three Essays on the Supply of Labor and the Demand for Goods," unpublished doctoral dissertation, Princeton Univ. 1971.

J. Hicks, *Capital and Growth*, Oxford 1965.

M. Hurd, "Changes in Wage Rate between 1959 and 1967," *Rev. Econ. Statist.*, May 1971, *53*, 189–99.

H. G. Lewis, "Hours of Work and Hours of Leisure," *Proceedings of the Industrial Relations Research Association*, 1957, 196–207.

R. Lucas and L. Rapping, "Real Wages, Employment, and Inflation," in E. S. Phelps, ed., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

F. Modigliani and R. Brumberg, "Utility Analysis and the Consumption Function: An Interpretation of Cross Section Data," in K. Kurihara, ed., *Post Keynesian Economics*, New Brunswick 1954.

K. Nagatani, "Life Cycle Saving: Theory and Fact," *Amer. Econ. Rev.*, June 1972, *62*, 344–53.

R. Pfouts, "Hours of Work, Savings, and the Utility Function," in R. Pfouts, ed., *Essays in Economics and Econometrics*, Chapel Hill 1960.

F. Ramsey, "A Mathematical Theory of Saving," *Econ. J.*, Dec. 1928, *38*, 543–59.

P. Samuelson, *Foundations of Economic Analysis*, Cambridge 1947.

L. Thurow, "The Optimum Lifetime Distribution of Consumption Expenditures," *Amer. Econ. Rev.*, June 1969, *59*, 324–30.

J. Tobin, "Life Cycle Saving and Balanced Growth," in W. Fellner et al., eds., *Ten Economic Studies in the Tradition of Irving Fisher*, New York 1967.

M. Yaari, "On Consumer's Lifetime Allocation Process," *Int. Econ. Rev.*, Oct. 1964, *5*, 304–17.

Money Illusion and the Aggregate Consumption Function: Note

By GEORGE D. CRAIG*

This paper presents empirical evidence which strongly supports the findings of William Branson and Alvin Klevorick which appeared in the December 1969 issue of this *Review*. The results given below are based upon an aggregated price variable which should be disaggregated in further work *if* the multicollinearity problem can be overcome. That is, Alex Cukierman's suggestion (in the March 1972 *Review*) should be followed if it can be implemented without introducing so much collinearity that it biases the results towards concluding that there is no money illusion. It should be pointed out that even with a great deal of multicollinearity, Cukierman empirically accepts the same conclusion that Branson and Klevorick accept. They all find money illusion to be present. We further note that this conclusion is with reference to "actual" consumption, which excludes consumer expenditures on durables (but includes depreciation and imputed rent for consumer durables, as well as consumer expenditures on nondurables and services).

I. Results of the Estimation of Four Aggregate Consumption Functions

The following results extend the findings of Branson and Klevorick in several ways: (a) predictive accuracy is added as a criterion in order to help differentiate among functions; (b) data are quarterly from 1952-I to 1970-I instead of 1953-I to 1965-IV; (c) the functions are estimated with and without the price variable so that predictive accuracy can be compared; and (d) two wealth variables are used instead of just one.

The four functions are, without lagging (in the Branson-Klevorick notation):

$$c_t = e^{\beta_0} y_t^{\gamma} l_t^{\delta} e^{\epsilon_t} \tag{1}$$

$$c_t = e^{\beta_0} y_t^{\gamma} w_t^{\delta} e^{\epsilon_t} \tag{2}$$

$$c_t = e^{\beta_0} y_t^{\gamma} l_t^{\delta} P_t^{\eta} e^{\epsilon_t} \tag{3}$$

$$c_t = e^{\beta_0} y_t^{\gamma} w_t^{\delta} P_t^{\eta} e^{\epsilon_t} \tag{4}$$

(Branson-Klevorick function)

Functions 3 and 4 are exactly the same as functions 1 and 2, respectively, except that P_t^{η} enters multiplicatively into each function. Functions 3 and 4 will be called the P_t functions in what follows. All variables[1] except P_t, are real per capita; c_t is actual consumption (defined above) in period t; y_t is disposable personal income in period t; l_t, and w_t are, respectively, liquid assets and net worth for households at the beginning of period t; and P_t is the all-items *CPI* in period t.

One difference from Branson and Klevorick's study is that disposable personal income is used here in place of labor income. This is done for two main reasons (besides the fact that labor income is particularly difficult to derive accurately and currently on a quarterly basis): 1) Branson and Klevorick found that their main conclusions were not sensitive to this substitution, and 2) comparison to other consumption functions is facilitated if the same income variable is used. Point 1) is strongly supported by our quantitative results in Table 1, where function (4) is the modified Branson-Klevorick function. It is seen that the coefficients on *ln* y_t and *ln* w_t (.662 and .113) are very close to those found by Branson and Klevorick (.661 and .127).

The motivation for using the second wealth variable, liquid assets, is to see how sensitive the results are to the selection of the

---

* Associate professor of finance, Northern Illinois University. I would like to thank William Brown and Paul Eaton for their helpful comments. I am also indebted to the referees for many useful suggestions.

[1] See the Appendix for further details.

TABLE 1—ESTIMATED COEFFICIENTS FOR THE FOUR FUNCTIONS[a]

Function	$ln\ y_t$	$ln\ l_t$	$ln\ w_t$	$ln\ P_t$	b_0
(1)	1.745 (13.97)	−0.590 (−6.01)			−1.403 (−5.95)
(2)	0.730 (37.13)		0.258 (14.58)		−0.385 (−7.51)
(3)	0.858 (30.77)	−0.134 (−5.16)		0.427 (19.24)	b
(4)	0.662 (47.52)		0.113 (7.51)	0.312 (14.87)	b

[a] T-values are in parentheses.

[b] The intercept (b_0) is not significant ($\alpha = .05$) and is suppressed.

wealth variable. Since l_t is the most commonly used wealth variable in consumption functions, along with net worth, l_t is selected. (If the linear form were utilized, function (1) would be James Tobin's consumption function.) Of course, liquid assets are only a small portion (about 15 percent) of net worth and represent the most visible part of wealth for the majority of consumers. Thus the two variables may be related much differently to actual consumption. Whether this is true or not, it is of interest to notice any differences between the results when using w_t and l_t, particularly with respect to the money-illusion issue.

The predicting takes place at intervals of four quarters, starting with the prediction of 1963-I through 1963-IV by using the sample period of 1952-I through 1962-IV. The four predicted periods are then added to the sample period and the following four quarters (or less if data are nearly exhausted) are predicted. Thus the sample period begins with 44 quarters and continues to grow until it includes all 73 quarters. Therefore, 40 percent of the time period has been predicted (using reported values for the independent variables) with each of the four functions.

The results of the estimations are presented in Tables 1 and 2.[2] Functions (3) and (4) (the P_t functions) are the consumption functions which either indicate the presence or the absence of money illusion (called money-illusion consumption functions in the Branson-Klevorick terminology). Each variable used in functions (1) through (4) is transformed into *logs* and is lagged as specified in Table 2. As an example, the logarithmic form of function (4) with lagging is:

$$ln\ c_t = \beta_0 + \sum_{i=0}^{I} \gamma_i\, ln\ y_{t-i} + \sum_{j=0}^{J} \delta_j\, ln\ w_{t-j} + \sum_{k=0}^{K} \eta_k\, ln\ P_{t-k} + \epsilon_t$$

Table 1 gives the regression results for the functions. The money-illusion consumption functions (3) and (4) can be compared directly to their counterparts (1) and (2) which do not allow any money-illusion effect, if present, to surface. In terms of the coefficients for $ln\ P_t$ and their respective t-values, the results in the P_t functions do not differ very much. They are both fairly similar to the Branson-Klevorick coefficient (.418) and both indicate the presence of money illusion. Thus the presence of money illusion is not sensitive to the choice of the wealth variable here, even though the effects on $ln\ c_t$ are of opposite sign. Note that in function (4) the t-value for $ln\ P_t$ is even larger than that found by Branson and Klevorick (11.6) when they used the shorter period of 1953-I through 1965-IV.

A few more points worth noting in Table 1. When liquid assets are used the coefficient on $ln\ y_t$ is much more reasonable in the P_t function (3) than when P_t is not included (function (1)). Furthermore, the t-values for $ln\ y_t$ increase substantially in both cases when P_t is included. Lastly, the influence of the wealth variables is decreased greatly when we move to the P_t functions. Since the absence of autocorrelation is indicated in the P_t functions but not in functions (1) and (2) (see Table 2), the coefficients in the P_t functions are more reliable. All of these points indicate that the inclusion of the price variable is quite worthwhile.

The results in Table 2 support the Branson-Klevorick approach, as well as the specific formulation of their money-illusion consumption function (4). The first point to

---

[2] The results in Tables 1 and 2 are based upon all 73 quarters, and they are all from the estimation of the logarithmic form exemplified below, with lags.

TABLE 2—RESULTS OF THE ESTIMATION OF THE FOUR FUNCTIONS

Function	Average Relative Error (percent)	$\sum_{k=0}^{K} \eta_k$	δ_0	Short-Run Marginal Propensity to Consume[d]	*D.W.*[e]	$\bar{R}^2$ [f]	Lag Specifications[g]
(1)	1.516		−0.116[c]	0.482	0.532 (+)	.9957	$y_t\,(9,3), l_t\,(9,3)$
(2)	0.493		0.052[c]	0.279	0.978 (+)	.9984	$y_t\,(9,3), w_t\,(9,3)$
(3)	*0.407*[a]	0.427[b] (19.24)	0.026[c]	0.278	1.614 (No)	.9991	$y_t\,(10,2), l_t\,(9,2), P_t\,(10,2)$
(4)	*0.278*[a]	0.312[b] (14.87)	0.113[c]	0.180	2.068 (No)	.9994	$y_t\,(10,2), w_t\,(4,3), P_t\,(10,2)$

[a] *Italic* indicates that the function predicts better than both of the naive models.

[b] The *t*-value for the sum of the price coefficients is in parenthesis.

[c] Elasticities of c_t with respect to l_t or w_t which are based upon the present quarter effect only.

[d] SRMPC $= (\gamma_0)\,(c_{1970-I}/y_{1970-I})$ where $\gamma_0 = (\partial c/c)/(\partial y/y)$ for the current period.

[e] Results of the Durbin-Watson test: + (No) indicates positive (No) autocorrelation. A two-tailed test is used with a .05 level of significance. All *D.W.* results are from the nonhomogeneous forms.

[f] Corrected for degrees of freedom.

[g] First number is the number of quarters lagged. The second number is the degree of the polynomial.

note in Table 2 is that the average relative error (*ARE*) of the predictions are both reduced when the price variable enters into each of the functions (from 1.5 to 0.4 percent and from 0.5 to 0.3 percent).[3] Furthermore, functions (1) and (2) don't even predict as well as the best of the naive models (no-change and equal-change), whereas both of the P_t functions predict better than the best naive model.[4]

Continuing across Table 2, having already discussed the coefficients on *ln* P_t above, it can be seen that the inclusion of P_t brings the values of the short-run marginal propensity to consume into a smaller range: .18 to .28 with P_t as opposed to .28 to .48 without P_t. Next, there is no autocorrelation indicated in the P_t functions whereas positive autocorrelation is indicated in functions (1) and (2). Thus we can have more confidence in the results of the P_t functions, particularly in the significance indicated for each variable. The $\bar{R}^2$s are both higher for the P_t functions, and, in particular, it is best for function (4). The lag specifications are discussed below in detail.

The outstanding point in Table 2 is that function (4), the Branson-Klevorick money-illusion consumption formulation, predicts best, fits best, has no autocorrelation indicated, and gives very reasonable coefficient values. Thus, the Branson and Klevorick function performs very well when the time period covered has been increased by 21 quarters.

The lag specifications in Table 2 need to be explained before we further compare our specific results for function (4) with the Branson-Klevorick results. Following Shirley Almon's suggestion, p. 181, we first run each independent variable against *ln* c_t, lagging the independent variable up to 12 quarters and using second and then third degree polynomials. Judging the results mostly by $\bar{r}^2$ we get first approximations for lag length and polynomial degree. Then the multiple regressions are run for functions (1) through (4). Based on the first approximations, various combinations are used for the lag lengths and the two different polynomial degrees. The criteria used to determine the best combination are maximum $\bar{R}^2$ and the maximum *t*-

---

[3] Average relative error = (sum of the percents of predictive error)/(number of predictions). The root-mean of the squared-errors of the predictions are also all reduced when the price variable enters into each of the functions.

[4] The two naive models are $C_t = b_0 + b_1 C_{t-1}$ (no-change model) and $(\Delta C)_t = b_0 + b_1 (\Delta C)_{t-1}$ (equal-change model). Each model is run with b_0 suppressed (equal to zero) and with it unsuppressed. The best *ARE* value among the four runs is with the no-change model with b_0 suppressed (i.e., $C_t = b_1 C_{t-1}$), where the *ARE* is equal to .454 percent.

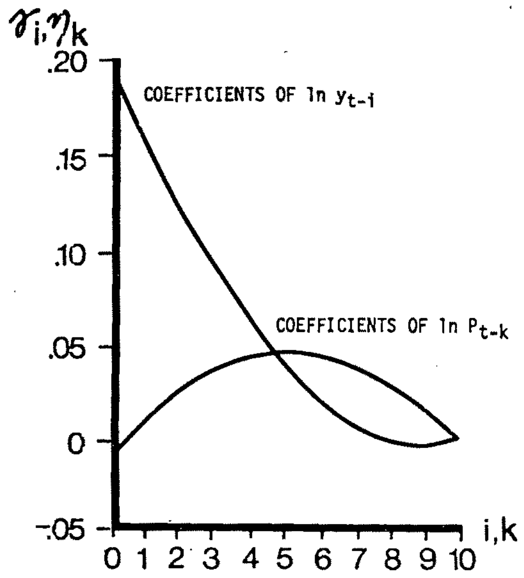


FIGURE 1. DISTRIBUTION OF COEFFICIENTS OF $\ln y_{t-i}$ AND $\ln P_{t-k}$ FOR FUNCTION 4

value for the sum of the coefficients for each lagged independent variable. If the variable's sum coefficient is not more significant than the variable's coefficient without lagging, then the variable is not lagged. The lags used, based upon the above procedure, are given in Table 2. Following the Branson and Klevorick suggestion, we keep the length of lagging on $\ln y_t$ and $\ln P_t$ equal to each other. In function (4), $\ln w_t$ is lagged slightly but is not much more significant than its unlagged form.[5]

The lagging schemes for $\ln y_t$ and $\ln P_t$ in function (4) are given in Figure 1.[6] Our distribution of the coefficients for $\ln y_t$ is very similar to the Branson-Klevorick distribution, p. 840, except that our lag length is ten instead of seven. The distribution of coefficients for $\ln P_t$ in Figure 1 has a different shape than the final distribution used by Branson and Klevorick. However, they are not very sure about the exact shape of the $\ln P_t$ distribution, p. 841. In fact, they do get at least one result which is shaped similarly to our distribution (see their Figure 1, p. 840). Our P_t results are similar to theirs in that they both are not significantly different from zero in the current period. The t-values for our $\ln P_t$ coefficients are -0.28 and 1.57 for the current quarter and the preceding quarter, respectively. (The critical t-value for a right tail test at the .05 level is 1.67.) All of our t-values for $\ln P_t$ are greater than 4.50 after the first two values. Our $\ln P_t$ coefficients are significant in 8 of the 10 lags whereas the Branson and Klevorick results are significant in only 3 of the 7 lags.

Using the same significance test with y_t as we use with P_t we find that the $\ln y_t$ coefficients are significant in all periods except the earliest two periods (8 and 9). (Each t-value is greater than 5.00 back through the sixth previous quarter, and it is 1.73 in the seventh previous period.) The reason that $\ln y_t$ is not cut off at the eighth period is that the coefficients for $\ln P_t$ show up well at least through the ninth previous period, as indicated above. Setting them both at a lag length of ten is a middle ground compromise.

Hence, function (4) differs from the Branson-Klevorick function in two ways: 1) substitution of disposable personal income for labor income, and 2) a different shape for the lag distribution of coefficients on P_t. Branson and Klevorick themselves find that the results are not very sensitive to either change, pp. 841–42. Thus we consider function (4) to be a good approximation to the Branson-Klevorick formulation.

Beyond the results of Table 2, we have found in a much broader investigation of consumption functions that function (4) still is best. That is, the Branson-Klevorick function performs best in terms of *ARE*, *D.W.*, $\bar{R}^2$, and the significance of the coefficients.

---

[5] In three of the four functions, a lag of 9 quarters is found to be best with the wealth variables in terms of our criteria. This does make sense, particularly with respect to l_t since this may be the average length of time needed to accumulate liquid assets in order to buy, or make down payments on, consumer durables. The negative value for the sum of the coefficients for l_t says that the amassing of liquid assets has a negative effect upon expenditures on nondurables and services, along with the depreciation and imputed rent on consumer durables. This result can be explained but the presentation of the full analysis is beyond this paper.

[6] As can be seen in Figure 1, the coefficient is forced to be zero in the earliest lagged quarter. The distribution is unconstrained in the current period. Both of these approaches are the same as those used by Branson and Klevorick.

The broader testing is done using both *log* and linear forms of most of the prevalent consumption functions (complements to the saving functions presented in my 1970 paper, p. 1132). Both *log* and linear functions are estimated with and without Almon lags, as well as with and without the price variable when l_t or w_t is included in a function. The presentation of all of these results is, of course, beyond the scope of this present paper.

II. Conclusions

The above analysis and results point to the conclusion that the consumption function which Branson and Klevorick put forth is a very useful formulation and performs well over the extended period of 1952-I to 1970-I. Beyond the implication of the presence of money illusion, there is a strong indication to include a separate price variable, when a wealth variable is used anyway. The latter statement is especially true if one is interested in maximizing predictive accuracy, which is an indicator of the correctness of specification.

APPENDIX

Description of the Data

In the estimation of the regressions, 9 to 10 of the initial quarters are lost, depending upon the maximum lag used in the equation. The number of quarters lost is equal to the number of quarters actually used with nonzero coefficients.

The following tabulation describes the data used in this study.

Actual consumption (C_t)	Billions of $	FRB-MIT-PENN model
CPI: all items (P_t)	1957–1959 = 100	*Surv. Curr. Bus.*
Disposable personal income (Y_t)	Billions of $	Same
Household liquid assets (L_t)	Same	*Fed. Reserve Bull.*, flow-of-funds data
Household net worth (W_t)	Same	FRB-MIT-PENN model
Population (N_t)	Millions	*Curr. Pop. Reports*, Series P-25
Price deflator for actual consumption (*PCON*)	1958 = 100	FRB-MIT-PENN model

Notes: The *CPI* data are quarterly figures derived by averaging three months. Liquid assets are household holdings of currency, demand deposits, U.S. government savings bonds, and savings and time deposits. Population data are averages of the last two months in each quarter. In the study, C_t, L_t, W_t, and Y_t are deflated by *PCON* and by N_t.

REFERENCES

S. Almon, "The Distributed Lag Between Capital Appropriations and Expenditures," *Econometrica*, Jan. 1965, *33*, 178–96.

W. H. Branson and A. K. Klevorick, "Money Illusion and the Aggregate Consumption Function," *Amer. Econ. Rev.*, Dec. 1969, *59*, 832–49

G. D. Craig, "Predictive Accuracy of Quarterly and Annual Aggregate Saving Functions," *J. Amer. Statist. Assn.*, Sept. 1970, *65*, 1131–45.

A. Cukierman, "Money Illusion and the Aggregate Consumption Function: Comment," *Amer. Econ. Rev.*, Mar. 1972, *62*, 198–206.

J. Tobin, "Asset Holdings and Spending Decisions," *Amer. Econ. Rev. Proc.*, May 1952, *42*, 109–23.

Semivariance and Stochastic Dominance: A Comparison

By R. Burr Porter*

Recent papers have examined theoretical and empirical conflicts between stochastic dominance *SD* and mean-variance *EV* decision rules. The purpose of this paper is to provide a corresponding analysis of relationships between *SD* and mean-semivariance *ES* criteria. Section I briefly summarizes the results of prior comparisons of *EV* and *SD*. Section II examines the concept of *ES* efficiency and shows that a priori we should expect greater empirical consistency between *ES* and *SD* than between *EV* and *SD* efficiency.

I. Results of *EV* vs. *SD* Tests

It is well known that when empirical distributions are nonnormal there exists the potential for serious conflicts between choices made by the *EV* rule and those made by one of the *SD* rules. In developing the first *SD* rule (*FSD*), James Quirk and Rubin Saposnik produced such an example using uniform density functions. On the other hand, several factors serve to limit the extent of *EV* vs. *SD* conflicts. For example, in studying the second *SD* rule (*SSD*), my 1971 paper has shown that:

1. with uniform density functions, if one is efficient by *SSD* it is efficient by *EV*;
2. with exponential density functions, the *SSD* efficient set is a proper subset of the *EV* efficient set;
3. with *gamma* density functions, the *EV* and *SSD* rules can make opposing choices only in a restricted subset of cases;
4. regardless of the form of the probability function, if two functions have the same mean (variance) but different variance (mean) and one is inefficient by *SSD*, then it also is inefficient by *EV*.

While these results may limit potential conflicts between *EV* and *SSD* choices, the true significance of such differences is an empirical question. Recent studies of 138 individual securities from the Israel Stock Exchange by Haim Levy and Giora Hanoch and of 149 mutual funds by Levy and Marshall Sarna concluded that the *EV* and *SSD* efficient sets were remarkably similar (but definitely not identical) in both size and composition. More extensive tests by Jack Gaumnitz and myself involving over 900 portfolios of securities randomly selected from the Chicago Price Relative File provided some support for the idea that *EV* and *SSD* choices are empirically similar. But we did note some differences—especially the tendency for the *EV* efficient set to include some low return, low variance portfolios that were eliminated by the *SSD* rule—that could be significant to a risk averse investor. A different type of study by Maurice Joy and myself employed the *SSD* rule as a test of mutual fund performance and obtained results generally consistent with those produced by William Sharpe using an *EV* test.

In summary, the evidence seems to suggest that actual conflicts between *EV* and *SSD* efficiency occur with significant frequency, but that the seriousness of the disagreements has yet to be fully evaluated. In the meantime, there is good reason to suspect that the use of an *ES* rule instead of the *EV* rule may lead to choices that are much more consistent with the conceptually superior *SSD* rule.

II. The *ES* Concept of Portfolio Efficiency and Its Relationship to *SD* Efficiency

The definitions of *ES* dominance and efficiency are the same as those of the corresponding *EV* relationships with semivariance *S* substituted for variance *V*. That is, given two portfolios, *A* and *B*, *A* dominates *B* by the *ES* rule if either

---

* Associate professor of finance, Pennsylvania State University. Much of the work in preparation for this paper was completed at the University of Kansas. I wish to acknowledge the helpful comments of O. Maurice Joy and an anonymous reviewer.

$$\text{(1)} \quad \text{a)}\ E_A > E_B \ \text{ and } \ S_A \leq S_B,$$

or

$$\text{b)}\ E_A \geq E_B \ \text{ and } \ S_A < S_B$$

Any portfolio that is not dominated by any other portfolio by this rule is a member of the *ES* efficient set.

As suggested by Harry Markowitz and developed more fully by James Mao (1970b), the *ES* rule can be applied in either of two distinctly different ways. One approach is to measure the semivariance S_E as the expected value of deviations below the mean. Following this method, S_E is defined as

$$\text{(2)} \quad S_E = \begin{cases} E[(R - E(R))^2] & \text{for } R < E(R) \\ 0 & \text{for } R \geq E(R) \end{cases}$$

where R is the random variable. The reference point, E, is different for each portfolio.

A second version measures semivariance as the expected value of deviations below a critical (or target) value If we let h represent the fixed reference point, then

$$\text{(3)} \quad S_h = \begin{cases} E[(R - h)^2] & \text{for } R < h \\ 0 & \text{for } R \geq h \end{cases}$$

The appeal of S_h as a measure of risk is that it is consistent with the practitioner's view of risk as reported by R. F. Lanzilotti and by Mao (1970a).They found that executives consistently define or explain risk as the chance of failing to meet their target level of return. In contrast to the *EV* rule that considers deviations from E to be equally undesirable whether they are in the direction of greater return or smaller return, ES_h penalizes only deviations below the target return.

Another basis for acceptance of the ES_h rule is the fact that the use of ES_h is consistent with maximization of expected utility where the utility function has the form

$$\text{(4)} \quad U(R) = \begin{cases} a+bR+c[(R-h)^2] & \text{for } R<h \\ a+bR & \text{for } R\geq h \end{cases}$$

where a, b, and c are constants, $b>0$ and $c<0$.[1] It is important to note the difference between this utility function—which is quadratic only for $R<h$—and that frequently associated with the *EV* model. The first derivative of the *EV* utility function eventually becomes negative as R increases beyond some point. It is, therefore, inconsistent with the requirements for *FSD* and *SSD*, one fact that helps account for the conflicts between *EV* and *SSD* efficiency reported in the literature reviewed above. Since the ES_h utility function is monotone increasing, it does satisfy the requirements of *FSD*. Therefore, the ES_h efficient set is a subset of the *FSD* efficient set. Moreover, the fact that the second derivative of the utility function is negative for $R<h$ indicates that it is consistent with the requirements of *SSD* over a significant portion of the domain of its argument. In fact, as h becomes very large, the ES_h efficient set approaches a subset of the *SSD* efficient set.[2] These observations suggest that the relationship between ES_h and *SSD* efficiency may be quite close. We support this suggestion by proving that except for the possible case of ES_h efficient portfolios with identical mean and semivariance, the ES_h efficient set is a subset of the *SSD* efficient set for any value of h.[3]

THEOREM: *Except for portfolios with identical mean and semivariance, every portfolio that is efficient by the ES_h criterion also is efficient by the SSD criterion.*

Discussion:[4] If every portfolio that is efficient by the ES_h criterion also is efficient by the *SSD* rule, then the ES_h efficient set is contained in—or is a subset of—the *SSD* efficient set. We prove that such is the case by showing that if a portfolio F dominates a portfolio G by *SSD* then it must be true that F dominates G by ES_h. This result then implies that if G is efficient—that is, is not dominated by any other portfolio—by the ES_h rule, G must be efficient by the SSD rule. Note that since it is not true that SSD efficiency implies ES_h efficiency, the ES_h efficient set frequently will be a proper subset of the SSD efficient set.

---

[1] See Markowitz and Mao (1970b) for proofs of the relationship between ES_h and utility functions of this form.

[2] It is important to note that these results do not hold for the comparison of ES_E and *SD* efficiency since ES_E implies a much more complex utility function containing a structural parameter that is a function of the means of the random variables.

[3] This result also fails to hold for ES_E versus *SD* efficiency. It is quite easy to construct portfolios that are inefficient by any *SD* rule but efficient by ES_E.

[4] The discussion and proof assume that no two portfolios that are efficient by ES_h have identical E and S_h values.

PROOF:

1. If F dominates G by SSD then

$$\int_{-\infty}^{a} [F(x) - G(x)]dx \leq 0 \tag{5}$$

for all values of a, with strict inequality for at least one value of a, where $F(x)$ and $G(x)$ are the distribution functions of return on portfolios F and G, respectively.

2. By definition, the semivariance for F and G can be written as

$$S_h(F) = \int_{-\infty}^{h} (x - h)^2 dF(x) \tag{6}$$

$$S_h(G) = \int_{-\infty}^{h} (x - h)^2 dG(x) \tag{7}$$

Integrating by parts twice and simplifying[5] we have

$$S_h(F) = \int_{-\infty}^{h} \int_{-\infty}^{a} F(x)dxda \tag{8}$$

$$S_h(G) = \int_{-\infty}^{h} \int_{-\infty}^{a} G(x)dxda \tag{9}$$

Subtracting (9) from (8) we have

$$S_h(F) - S_h(G) = \int_{-\infty}^{h} \int_{-\infty}^{a} [F(x) - G(x)]dxda \tag{10}$$

From step 1, equation (5), we know that the inside integral and therefore the entire right side of (10) is nonpositive. This result implies that the semivariance of G is at least as large as the semivariance of F.

3. The fact that F dominates G by SSD implies that the mean of F is at least as large as that of G. This result together with the semivariance relationship derived in step 2 and our assumption that F and G differ in either mean or semivariance is sufficient to guarantee that F dominates G by the ES_h criterion. Since SSD dominance implies ES_h dominance, we conclude that with the one possible exception noted, the ES_h efficient set is a subset of the SSD efficient set.

III. An Empirical Comparison of ES and SSD Efficiency

Although we have considered several conceptual bases for expecting closer agreement between the ES_h and SSD efficient sets than between either the ES_E and SSD or the EV and SSD efficient sets, an empirical test of differing degrees of similarity seems desirable. In order to satisfy this requirement we have performed the various efficiency tests on a set of 893 portfolios of securities selected randomly from the Chicago Price Relative File.

For each portfolio, 72 individual monthly rates of return were determined and were employed in calculating the means, variances, semivariances, distribution functions and integrals of distribution functions necessary for the various efficiency tests. The algorithms and procedures employed in conducting the tests are those developed by Porter, James Wart, and Don Ferguson. Efficient sets were calculated according to five criteria: SSE, EV, ES_E, $ES_{0.0}$, and $ES_{0.1}$.[6]

The results, shown in Table 1, basically are consistent with those expected. The $ES_{0.0}$ set contains 50 members, all of which are included in the SSD set. The $ES_{0.1}$ set contains 49 members, all of which are included in the SSD set and 43 of which also are in the $ES_{0.0}$ set. In comparison, the EV efficient set includes 69 members, 8 of which are excluded by the SSD rule. The one surprise is that all of the 70 members of the ES_E set also are members of the SSD set. Our discussion in Section II above anticipated contradictions between ES_E and SSD that

---

[5] Let $U=(x-h)^2$ and $dV=f(x)dx$. Then let $U=(x-h)$ and $dV=F(x)dx$. The simplifications result from the facts that $F(-\infty)=0$ and that $(x-h)=0$ for $x=h$.

[6] The $ES_{0.1}$ test should be interpreted as semivariance around a 10 percent *annual* rate of return. A monthly rate of 0.8 percent was used as a simple approximation to an annual rate of 10 percent.

TABLE 1—SIZE OF THE EFFICIENT SET[a]

Type of Data	Efficiency Criteria				
	SSD	EV	ES_E	$ES_{0.0}$	$ES_{0.1}$
Monthly	216	69	70	50	49
Quarterly	127	65	62	33	33
	$ES_{0.0}$ and $ES_{0.1}$	SSD and EV	SSD and ES_E	SSD and $ES_{0.0}$	SSD and $ES_{0.1}$
Monthly	43	61	70	50	49
Quarterly	26	45	55	33	33

[a] The top half of the table gives the sizes of various individual efficient sets. The bottom half gives the number of portfolios common to several pairs of efficient sets. With monthly data, for example, 61 of the 69 members of the EV efficient set also included in the SSD efficient set.

failed to materialize in this sample. In addition, the ES_E set includes most of the members of the $ES_{0.0}$ and $ES_{0.1}$ sets.

As a further check on the consistency of the different efficiency criteria the monthly observations for each of the 893 portfolios were converted to quarterly rates of return, and the efficiency tests were repeated. These results, also shown in Table 1, are quite similar to the results using monthly data. The major differences are that all of the efficient sets are smaller than before[7] and the ES_E set contains 7 members (out of a total of 62) that are not efficient by the SSD rule. With quarterly data, therefore, we did find some difference between ES_E and SSD efficiency, but this discrepancy still is not as significant as that between EV and SSD in which 20 out of 65 EV efficient portfolios are dominated by the SSD rule.

IV. Summary and Implications

This paper contains two main results. First, there are good conceptual bases for anticipating greater consistency between ES and SSD rules than between EV and SSD rules. In particular, with one possible and relatively minor exception (identical mean and semivariance), the efficient set of portfolios derived from the ES rule with semivariance calculated around a fixed point is a subset of the SSD efficient set. Second, our empirical results confirm our expectations. The ES efficient sets with semivariance around a fixed point are subsets of the SSD set including less than one-quarter as many members as the SSD set. The ES efficient sets with semivariance around the mean show much more agreement with the SSD efficient sets than do the EV efficient sets.

In evaluating the various efficiency criteria, it seems fair to say that attention should be focused more on SSD and ES with semivariance around a fixed point. The EV rule implies violations of the assumptions of rational behavior and leads to decisions that are empirically unjustified. The use of semivariance around the mean as a measure of risk also leads to incorrect choices, but with less frequency. With semivariance around a fixed point as the measure of risk, the ES rule will not accept portfolios that are eliminated by the SSD rule. Moreover, it appears to eliminate a substantial portion of the SSD efficient set. If, therefore, the decision maker is willing to choose on the basis of ES with semivariance around a fixed point, he can draw comfort from the fact that he is choosing from a subset of the SSD efficient set.

REFERENCES

J. Hadar and W. R. Russell, "Rules for Ordering Uncertain Prospects," *Amer. Econ. Rev.*, Mar. 1969, *59*, 25–34.

G. Hanoch and H. Levy, "The Efficiency Analysis of Choices Involving Risk," *Rev. Econ. Stud.*, July 1969, *36*, 335–46.

O. M. Joy and R. B. Porter, "Stochastic Dominance and Mutual Fund Performance," *J. Finance Quant. Anal.*, Jan. 1974.

R. F. Lanzilotti, "Pricing Objectives in Large Companies," *Amer. Econ. Rev.*, Dec. 1958, *48*, 921–40.

H. Levy and M. Sarnat, "Alternative Efficiency Criteria: An Empirical Analysis," *J. Finance*, Dec. 1970, *25*, 1153–58.

J. C. T. Mao, (1970a) "Survey of Capital

---

[7] I noted a similar result (1973) in working with EV and SSD efficient sets using different types of data. The fewer the data points the smaller the efficient set regardless of the efficiency criterion employed.

Budgeting: Theory and Practice," *J. Finance*, May 1970, *25*, 349–60.

———, (1970b) "Models of Capital Budgeting, E-V vs E-S," *J. Finance Quant. Anal.*, Jan. 1970, *4*, 657–75.

H. M. Markowitz, *Portfolio Selection*, New York 1959.

R. B. Porter, "A Comparison of Stochastic Dominance and Mean-Variance Portfolio Models," presented at the national meeting of the Financial Management Association, Oct. 1971.

———, "An Empirical Comparison of Stochastic Dominance and Mean-Variance Portfolio Choice Criteria," *J. Finance Quant. Anal.*, Sept. 1973, *8*, 587–608.

———, **and J. E. Gaumnitz**, "Stochastic Dominance vs. Mean-Variance Portfolio Analysis," *Amer. Econ. Rev.*, June 1972, *62*, 438–46.

———, **J. R. Wart, and D. L. Ferguson**, 'Efficient Algorithms for Conducting Stochastic Dominance Tests on Large Numbers of Portfolios," *J. Finance Quant. Anal.*, Jan. 1973, *8*, 71–83.

J. P. Quirk and R. Saposnik, "Admissibility and Measurable Utlity Functions," *Rev. Econ. Stud.*, Feb. 1962, *29*, 140–46.

W. F. Sharpe, "Mutual Fund Performance," *J. Bus., Univ. Chicago*, Jan. 1966, *39*, Part II, 119–38.

The Precautionary Demand for Reserves

By ERNST BALTENSPERGER*

The square root law of the precautionary demand for cash, or reserves of other assets, was known as early as 1888, when Edgeworth first mentioned it. Recently, J. H. G. Olivera presented a useful review and unification of the literature on this subject, including a review of the stochastic processes compatible with the result.[1] The square root rule implies, of course, a value of one-half for the elasticity of the demand for precautionary reserves with respect to scale (provided that an increase in scale means an increase in the number of transactions, rather than just an increase in the average size of transactions). Other models of precautionary demand for cash, such as those by Edward Whalen or Merton Miller and Daniel Orr suggest, due to somewhat different assumptions, an elasticity of one-third, i.e., even less than the value implied by the Edgeworth model (subject to the same provision). The term precautionary reserve is used in this paper to denote the excess of total holdings of the reserve asset in question over the expected or average net demand for ("loss" of) reserves; that is, in the same sense as Olivera uses it.

The source of the demand for precautionary reserves is, of course, the existence of uncertainty about net demand for reserves, the fact that reserve changes are known in a probabilistic form only. In the standard precautionary demand models, the optimal reserve is derived for a given probability distribution of reserve changes. That is, the degree of uncertainty normally measured by the variance of reserve changes is given—at least relative to a scale variable, such as the number or volume of transactions. It can be argued, however, that in many instances the degree of "knowledge" or "certainty" about reserve changes is, to a certain extent, subject to the control of the decision maker. An individual, or an enterprise, can spend time and other resources in order to gain information about the net demand for reserves they may have to face in a given time period. They can plan the nature and timing of their transactions (their sales and purchases) more or less exactly, make more or less efforts to collect information about prices and opportunities. To plan one's activities 100 percent would reduce uncertainty to a minimum, but it is likely to be prohibitively expensive, in terms of planning and information costs. To spend no resources at all on planning (thinking about the future) may, on the other hand, also be prohibitively expensive, in terms of the risk of surprises or emergencies, and the corresponding costs of adjustment. It seems likely that, in general, neither of these extremes would be economically efficient. Using resources for planning one's activities reduces uncertainty, the risk of surprises, and the associated costs of adjustment. Holding inventories of safety reserves is another method of reducing these costs. One of the main functions of precautionary reserves is, therefore, to save planning and information costs.

It is the purpose of this paper to show that, if the precautionary demand model is modified to take into account the economics of planning and information, the elasticity of the demand for precautionary reserves with respect to scale may be quite different from one-half, i.e., the value implied by the standard square root result. For example, it may well be in the neighborhood of one. This does not mean, of course, that the scale economies generated by the stochastic process underlying the model are nonexistent. However, they may manifest themselves in a different way and thus not appear so openly on the surface. The reason is that now substitution at two margins is possible. To increase inventories of reserve assets is one way of reducing the risk of reserve deficiencies and the associated adjustment costs. More planning and information is another way of achieving the same goal. The scale economies implied by the stochastic character of the model—summarized in the square root result—change the marginal gain from additional planning, and thus the optimal amount of time and other resources allocated to that activity.

---

* Associate professor of economics, Ohio State University.

[1] In particular, he has shown that the law is compatible with serial correlation among reserve changes.

The paper thus gives a model of the demand for precautionary reserves in which the degree of knowledge, or uncertainty, itself is a decision variable. The model does have obvious importance with regard to the issue of the transactions (scale) elasticity of the demand for precautionary reserves, and with respect to attempts at empirical verification of the precautionary demand model. It suggests that the scale economies implied by the stochastic character of the approach (manifested in the square root formula) may be "hidden" behind reductions in other cost elements, rather than appearing directly in the form of a relative decline of reserve holdings, as the standard formulation suggests.

I. The Model

Let X denote the net demand for reserves (net change or "loss" of reserves) over the planning period, and $\phi(X)$ the density function over X. Now, assume:

$$\sigma^2(X) = a^2T \tag{1}$$

where T is a scale variable (such as the number of transactions, which in turn may be a measure of the volume of transactions), and a is a parameter reflecting the degree of uncertainty (relative to the scale variable T). That is, we assume what Olivera, p. 1098, calls a constant structure of X, since this is the constellation which generates the square root result and the corresponding elasticity of one-half.[2] For simplicity, assume also that $\phi(X)$ is normal, and that $E(X)=0$.

Now let us summarize the argument in the preceding section by assuming that the parameter a is a function of the amount of resources per transaction (i.e., relative to the scale variable T) allocated to planning and gathering information, u. We hypothesize that an increase in u will reduce a, and that the marginal reduction of a will become smaller as u increases (diminishing marginal product of u; possibly, a cannot be reduced below some positive minimum level, which is approached asymptotically):

$$a = a(u) \tag{2}$$

with $\quad a'(u) < 0$

and $\quad a''(u) > 0$

As an example for the interpretation of this relationship, consider the case of Edgeworth's club manager (pp. 123–24). Presumably, he faces a population of club members who are not completely homogeneous as to their habits. But each of them may follow certain patterns, if only in a probabilistic sense. By learning about these, the manager can reduce the degree of uncertainty which he faces. Suppose, for example, that the members can be grouped into a number of different classes, and, to start with an extreme and simple case, that each class has a probability of either zero or one of appearing on a particular day of the week and ordering a particular item. Class one may have a probability of one of coming in on Mondays and ordering, say, steak, and a probability of zero of coming in on any other day of the week. Class two may always come in on Tuesdays and Thursdays, and so on. Under these circumstances, if the manager has complete knowledge of the habits of all members, he faces no uncertainty at all, and he does not need any precautionary reserves. But such a complete knowledge would require a certain effort. The manager would have to observe each members' habits over a period of time and maintain detailed records of them. If he does not make this effort, he will face uncertainty in the sense that he will observe variations of demand over time which he is not able to predict, which appear random to him (although they are, in principle, predictable—given sufficient information). By learning more about the members, he is able to reduce this uncertainty, and he will need less precautionary reserves. The argument does apply in the same way if probabilities (of appearing on a given day) between zero and one are allowed. Uncertainty can never be reduced to zero in this case, of course. But by learning about his customers, the manager will be able to differentiate better between predictable (expected) and unpredictable (random) variations of demand, and he will reduce uncertainty in this sense. Other situations can be described in similar terms, such as the case of a bank which holds safety reserves against default risk on loans, or the case of an insurance company.

---

[2] Of course, dropping assumption (1) will modify this result and lead to elasticities which differ from one-half. This is well known, and will thus not be taken up here. For example, if $\sigma(X)$, rather than $\sigma^2(X)$, is proportional to scale, we would get an elasticity of one. This might be the case if an increase in the total volume of transactions is brought about via increases in the size of each individual transaction, rather than the number of transactions possibly due to general price inflation. See Don Patinkin's discussion of this, pp. 87–88, and his reference to Karl Schlesinger.

Now, there are three kinds of costs which we have to take into consideration in our optimizing procedure:

First, the holding (or opportunity) cost of the reserve inventory (taken as a function of the beginning-of-period level of reserves).

Second, the adjustment cost occurring as a result of reserve deficiencies.

Third, the cost of planning and gathering information.

Let R denote the beginning-of-period inventory of the reserve asset in question, h the marginal holding cost (assumed constant, for simplicity), p the marginal adjustment cost per dollar of reserve deficiency (also assumed constant), u the input of resources per transaction (i.e., relative to T) into planning and gathering information, and w the price of these resources (for simplicity, consider just one resource input, say time, so that w would be the value of time). Then the expected total cost to be minimized, C, is:

$$(3) \quad C = hR + \int_R^\infty p(X-R)\phi(X)dX + wuT$$

It is convenient to rewrite this in terms of standardized units. Let $x = X/\sigma(X)$, and $b = R/\sigma(X)$.[3] Then we have:

$$\begin{aligned}(4) \quad C &= hb\sigma(X) + \sigma(X)\int_b^\infty p(x-b)f(x)dx \\ &\quad + wuT \\ &= \gamma\sigma(X) + wuT \\ &= \gamma a\sqrt{T} + wuT\end{aligned}$$

where

$$\gamma = \left[hb + \int_b^\infty p(x-b)f(x)dx\right]$$

and $f(x)$ denotes the standard normal distribution.

Optimization with respect to the decision variables b and u gives:

$$(5) \quad \frac{\partial C}{\partial b} = \left[h - p\int_b^\infty f(x)dx\right]a\sqrt{T} = 0$$

$$(6)^4 \quad \frac{\partial C}{\partial u} = \gamma\sqrt{T}\,a'(u) + wT = 0$$

Equation (5) implies that b must satisfy the following:

$$(5') \qquad \int_b^\infty f(x)dx = h/p$$

i.e., b (Olivera's "coefficient of security") is a parameter depending on h and p only, but not on T or u. Equation (6) implies that u must satisfy equation (6′).

$$(6') \qquad a'(u) = -\frac{w\sqrt{T}}{\gamma}$$

Given w and γ, an increase in the scale variable T will raise $a'(u)$ (in absolute terms), and thus, given our assumptions, will lower u or raise a.[5] This is so because of the scale economies caused by the stochastic nature of the underlying transactions. We can call the first term in (6) the marginal benefit from increasing u by a unit, and the second term the corresponding marginal cost. An increase in scale does lower the former relative to the latter (at any given level of u), and thus induces a substitution of holding and adjustment costs for planning and information costs (as compared to the situation when u is held constant).[6]

---

[3] Note that we did assume, for simplicity, $E(X)=0$.

[4] It can easily be seen that second-order conditions for a minimum are satisfied.

[5] Except in the case where $-a'(u)$ is smaller than $w\sqrt{T}/\gamma$ at any u, where we would get a corner solution with $u=0$. Note that γ is independent of T and u.

So we have

$$R = b\sigma(X) = ba\sqrt{T} \tag{7}$$

or

$$\log R = \log b + \log a + 1/2 \log T$$

where b is a parameter depending on h and p only (but not on T or u), and a is such that (6′) is satisfied.

From this, the elasticity of demand for precautionary reserves with respect to T follows as:

$$\frac{d \log R}{d \log T} = \frac{d \log a}{d \log T} + \frac{1}{2} \tag{8}$$

But as we just have seen, (2) and (6′) imply a positive elasticity of a with respect to T. An increase in T does raise the optimal level of $-a'(u)$, lower the optimal u, and raise a. Formally, note that from (2) we have $d \log a = [a'(u)/a(u)]\, du$, and from (6′) $d \log T = 2[a''(u)/a'(u)]\, du$, so that

$$\frac{d \log a}{d \log T} = \left[\frac{a'(u)}{a''(u)} \cdot \frac{a'(u)}{a(u)}\right]\left[\frac{1}{2}\right] > 0 \tag{9}$$

The first bracket term on the right-hand side is the elasticity of a with respect to $-a'$ and the second (one-half), the elasticity of $-a'$ with respect to T.

Thus the elasticity of optimal reserves with respect to scale, $d \log R/d \log T$, is more than one half given our assumptions. The exact amount by which it exceeds one-half depends on the form of the function $a(u)$,[7] and in fact the elasticity could well be one. It is easy to see that this would be the case for $a(u) = a_0 e^{-cu}$, which would give $a'(u) = -ca(u)$ and $a''(u) = c^2 a(u)$, so that $d\log a/d\log T = 1/2$ and $d\log R/d\log T = 1$.

II. Extensions and Qualifications

In the preceding discussion, we have assumed that the input of resources per transaction for planning and information (u) determines the degree of uncertainty per transaction ($a^2 = \sigma^2/T$). In other words, if total resources used for planning and information production ($uT \equiv I$) change in proportion to scale, total uncertainty σ^2 will also change in proportion to scale. That means, we assume no economies or diseconomies of scale in the production of information and planning. As long as the amount of time and other resources spent per transaction (per customer, etc.) is constant, a will stay constant. This does seem a neutral and, for our purposes, reasonable assumption.[8] But it does not have to be true, of course. We could start instead from a more general formulation relating total uncertainty to total resource input I and scale:

$$\sigma^2(X) = g(I, T) \tag{10}$$

with $g_I < 0$, $g_T > 0$

From this, we would obtain the case discussed above when the $g(I, T)$ function is linear homogeneous. An equiproportionate change in I and T—leaving the relative resource input u constant—would then change total uncertainty σ^2 in proportion. The formulation used in our preceding discussion does have the advantage, though, of conveniently separating the effects of scale and information on uncertainty, and thus is useful for the purpose of comparison with the standard precautionary demand model and the corresponding scale elasticity of one-half.

Under the conditions of our previous discussion, the marginal cost of lowering a by a given amount (the additional resource expense which must be incurred in order to reduce a by a given amount) is proportional to T. The marginal benefit from lowering a by a given amount (the reduction in the first two terms of the cost function (4) resulting from lowering a by a given amount), on the other hand, is proportional to $\sqrt{T}$. An increase in scale thus lowers the marginal benefit relative to the marginal cost, and hence our result that an increase in scale induces a relative decline of resources devoted to planning and information, and thus an increase in a. If, instead, the marginal cost of lowering a by a given amount (the additional resource expense required) were more than proportionate to T (i.e., if at a higher T we need relatively more resources to lower a by a given amount; in other words, if there are diseconomies of scale in planning and information production), an increase in T would lower the marginal benefit relative to the marginal cost even more than before, and thus the effect of T on the optimal a would be even stronger than described above. If, on the other hand, the marginal cost were less than proportionate to T (i.e., if at a higher T we need relatively fewer resources to lower a by a given amount; economies of scale in planning and information production), the effect of T on optimal a would be weaker than described above. However, as long as marginal cost depends on T with an elasticity exceeding one-half, the optimal a will increase in response to a rise in T.

---

[6] If these scale economies were nonexistent, we would not get this result. If $\sigma^2(X) = a^2 T^2$, i.e., if the standard deviation of X increases in proportion to T, instead of the square root of T, we would have $C = \gamma a T + wuT$, and $\partial C/\partial u = a'(u)\gamma T + wT = 0$, or $a'(u) = -\gamma/w$, independent of T.

[7] Of course, if a were a constant ($a'(u) = 0$, for all u), we would get the standard result of an elasticity of one-half.

[8] However, a fixed cost of producing any information, resulting in declining average costs, could be added without complication.

Another aspect of our previous discussion which requires some comment is the fact that we did consider a single planning period only. By doing so, we have implicitly assumed that planning and information expenses have to be repeated period after period if a given level of uncertainty is to be maintained. This may be appropriate with respect to certain planning and information activities, but not with respect to others. The model can be modified to apply to the case where resources spent on planning and information in the current period also affect the degree of knowledge in future periods. We then have to minimize the (discounted) value of all costs, present and future, affected by the current decision. That is, we minimize[9]

$$C = \sum_{t=0}^{\infty} \alpha^t \gamma \sigma_t + wuT \tag{11}$$

where α is a discount factor, $\sigma_t = a_t\sqrt{T}$ is the standard deviation of X in period t, and γ is as defined in (4).[10] The optimal u is found where

$$\frac{\partial C}{\partial u} = \sum_t \alpha^t \gamma \frac{\partial a_t}{\partial u} \sqrt{T} + wT = 0 \tag{12}$$

The first term represents the (discounted) marginal benefit of raising u by a unit,[11] and the second the corresponding marginal cost. An increase in T will again lower the former relative to the latter, and optimality requires that u is adjusted until the two are brought back to equality. If u affects all the a_t with diminishing returns (to the extent that it does affect them at all; for large enough t, the $\partial a_t/\partial u$ may become zero, of course), it is clear that this requires a reduction in u as T increases. Compared to the situation before (where $\partial a_t/\partial u = 0$ for all except the current period), the marginal value of planning and information is higher now, and more resources will—at any T—be allocated to these activities.[12] However, an increase in scale will still have the same effect of lowering the marginal benefit relative to the marginal cost, and thus induce a decline of u and an increase of a.[13]

---

[9] We neglect here costs of adjusting reserves to the optimum level.

[10] Since the optimal b and thus γ depend on h, p, and $f(x)$ only, they can easily be seen to be the same for all periods, determined as in (5′) (as long as no changes in h and p are anticipated). We can omit this part of the optimization procedure from explicit consideration here.

[11] In the extreme case where an information and planning expense made now lowers uncertainty by a given amount for all times ($\partial a_t/\partial u = a_u$, for all t>0), this term can be written as

$$\gamma a_u \sqrt{T} \sum_{t=0}^{\infty} \alpha^t = \gamma a_u \sqrt{T}/(1-\alpha)$$

[12] The current marginal expense in this case will, of course, exceed the current marginal gain, i.e., the decision maker will, at the margin, "take a loss" in the current period, in order to have the benefit of reduced uncertainty in future periods.

[13] Of course, a change in T might also change the levels of the resource inputs for planning and information which will be made in future periods, which in turn may affect the $\partial a_t/\partial u$ in equation (12). If we think in terms of an optimizing scheme where resource inputs for

Our major interest in this paper has been in the elasticity of the demand for precautionary reserves with respect to the scale variable T. However, it may be noted that the model does also have implications with regard to the elasticity of R with respect to the holding cost h (which may be interpreted as the interest elasticity in the case of the precautionary demand for cash). A change in h will require adjustments in the equilibrium levels of b and u (and thus a). Presumably, an increase in the holding cost h (raising the marginal cost of one way of reducing the risk of reserve deficiencies and the associated costs) would induce an increase in u (i.e., a more extensive use of an alternative way of reducing this risk), as well as a higher willingness to accept the possibility of such deficiencies and the corresponding costs (which are now relatively smaller at the margin). Both of these substitutions would help to explain a negative elasticity of R with respect to h.

Finally, the possibility should be noted that increases in income and transactions level (T) could be correlated with increases in adjustment costs p and resource costs w (for example, both of these contain the value of time as a component), which would tend to raise the elasticity of R with respect to T. However, important as this may be, it involves a different argument than the one presented in this paper. The results of this paper are obtained even if p and w are both independent of T.

III. Conclusions

A model of the precautionary demand for reserves has been presented which does incorporate planning and information activities and the associated costs. It has been shown that, in such a model, the elasticity of demand for precautionary reserves with respect to scale can be quite different from one-half (the value implied by the standard square root result) and may well be in the neighborhood of one. The scale economies implied by the stochastic character of the model can manifest themselves in ways other than a relative decline of reserve holdings.

---

planning and information and reserve holdings for many periods are planned simultaneously, the situation is somewhat more complex than described above, but our general result will still hold, under reasonable assumptions, at least in the sense that an increase in T raises the average level of a over the planning period.

REFERENCES

F. Y. Edgeworth, "The Mathematical Theory of Banking," *J. Royal Statist. Soc.*, Mar. 1888, *51*, 113–27.

M. H. Miller and D. Orr, "A Model of the Demand for Money by Firms," *Quart. J. Econ.*, Aug. 1966, *80*, 413–35.

J. H. G. Olivera, "The Square-Root Law of Precautionary Reserves," *J. Polit. Econ.*, Sept./Oct. 1971, *79*, 1095–04.

D. Patinkin, *Money, Interest and Prices*, 2d ed., New York 1965.

K. Schlesinger, *Theorie der Geld-und Kreditwirtschaft*, Munich 1914.

E. L. Whalen, "A Rationalization of the Precautionary Demand for Cash," *Quart. J. Econ.*, May 1966, *80*, 314–24.

Risk Premiums and the Life Cycle Hypothesis

By ROGER BOWDEN*

How does the consumer react to uncertainty in his future income receipts from human and nonhuman wealth? In answering this question, it is helpful to replace his stochastic life cycle problem with a certainty-equivalent problem in which expectations of future income receipts are adjusted by the subtraction of risk premiums, as a manifestation of risk aversion on the part of the individual.

A recent paper by Keizo Nagatani utilizes this idea. By supposing the consumer to maximize a lifetime utility function subject to an intertemporal budget constraint in which incomes are replaced by their certainty equivalents, he shows that with the unfolding of his years, the individual's planned consumption at any future date will shift up or down. Thus suppose the consumer is now aged v years and is planning consumption for age $t>v$. Then $C_v(t)$ will in general differ from $C_t(t)$.

However, the risk premiums on uncertain future income (from human wealth) that Nagatani uses are arbitrarily imposed, or exogenous to the maximizing problem. This seems wrong. Such premiums should reflect not only the variability of future incomes, but also the individual's reaction to such variability; or in other words, the nature of his intertemporal utility function.

Further, there is a prior question of uniqueness; are the risk premiums in each period uniquely defined? To see that this might prove potentially troublesome, imagine a decision problem:

$$\max_x Ef(x; \theta_1, \theta_2)$$

where θ_1 and θ_2 are random parameters with means μ_1 and μ_2. Under the certainty equivalent approach, one would write

$$Ef(x; \theta_1, \theta_2) = f(x; \mu_1 - \alpha_1, \mu_2 - \alpha_2)$$

where α_1 and α_2 are risk premiums which in general will depend on x. But since there is only one defining equation, any number of combinations of α_1 and α_2 satisfying this equation would be equivalent. This would greatly diminish the attraction of a supposition that the consumer behaves as if he had allowed for risk premiums closely tied to the distribution of θ_1 and θ_2, and the function f.

One can raise further questions. For instance, is it true (as Nagatani assumes) that the risk premiums attached to future incomes constitute a monotone-increasing function of time?

Such queries indicate if nothing else, that in order to be able to use risk premiums as an aid to understanding the consumer's life cycle decisions, we ought to know a great deal more about them. The present paper is organized towards that end. Specifically, I shall show:

(a) that risk premiums can indeed be uniquely defined;
(b) they are highly endogenous to the allocation problem, that is, they and the optimal consumption program are jointly determined;
(c) the risk premiums are not necessarily monotone-increasing. On the other hand it might, under reasonable assumptions, be possible to say something about their time profile. Finally,
(d) the implications of a plausible time profile of the premiums for manifest consumer behavior will be briefly explored.

Before we proceed any further, it is convenient at this point to recall some standard results in risk-aversion theory. Following John Pratt, we imagine a consumer with wealth x faced with an earning proposition z, a risk with mean μ and variance σ^2. The risk premium π is defined by:

$$(1) \qquad Eu(x + z) = u(x + \mu - \pi)$$

where u is the individual's utility function.

---

* Senior lecturer in economics, University of Auckland.

Thus the consumer would be indifferent between receiving the risk z and receiving the cash equivalent $E(z)-\pi$. We shall always assume $\pi>0$, so that the consumer is risk averse; in elliptical fashion we shall then refer to u as a risk-averse utility function.

Note that π will in general be a function of x. We shall usually be interested in the case where π is a decreasing function of x. For a small risk, one can make use of Pratt's measure of local risk aversion. Expanding the left-hand side of (1) to second order and the right-hand side to first order in Taylor series and taking expectations, one derives the approximation:

$$\pi \simeq \tfrac{1}{2}\sigma^2 \frac{u''(x+\mu)}{u'(x+\mu)}$$

Thus consider

$$u = \log x \quad \text{or} \quad u = \frac{1}{\gamma}x^{\gamma}, \qquad (0<\gamma<1)$$

In either case (and only in these cases) the measure of local risk aversion is proportional to $\sigma^2/x+\mu$. The individual's sensitivity to risk is inversely proportional to wealth. These are all the results we need.

I. The Lifetime Allocation Problem

It will be helpful to write our maximizing problem in dynamic programming format. We choose discrete and finite units of time, not only to avoid encumbering the exposition with mathematical detail, but also because income from human wealth is usually received in discrete lumps rather than as a continuous flow; we will therefore be able to distinguish between income expected in the coming period and that which actually accrued, from the vantage point of the end of the period.

We follow Nagatani in introducing uncertainty via a stochastic income receipt, which the consumer can spend or invest at a nonstochastic interest rate r. However, we do not allow borrowing, and in this respect our model has a choice aspect more like that of Edmund Phelps. The restriction on borrowing is not strictly necessary for the ensuing development, indeed I shall later say something about such a possibility; but it would bog down the exposition of some later points. We adopt a common dynamic programming time convention by subscripting variables with the number of periods to go until the end of the (certain) horizon. Thus, write x_N as the consumer's wealth with N periods to go. Then

$$x_{N-1} = r(x_N - C_N) + y_N$$

where y_N refers to income received during period N. This is unknown as of the start of the period. On the other hand, C_N is consumption planned at the start of the period and executed during the period.

We have now to say something about the subjective probability distribution of the incomes, y_r. It is reasonable to suppose that successive values of income should be correlated Given that there are N periods to go, the consumer's expectations are thus formed according to a scheme like:

$$E(y_r \mid y_{r+1}, y_{r+2} \ldots y_N) = f(y_{r+1}, y_{r+2} \ldots y_N)$$

For the sake of simplicity, we assume a Markov scheme:

$$E(y_r \mid y_{r+1}, y_{r+2} \ldots y_N) = \phi_r y_{r+1}; \qquad \phi_r \text{ constant}$$

By varying the ϕ_r, we can trace out different expected income profiles.

With regard to the conditional variances, we assume that

$$\mathrm{Var}\,(y_r \mid y_{r+1} \ldots y_N) = \sigma_r^2; \qquad \sigma_r \text{ constant}$$

Although it is not strictly necessary, it will again simplify the exposition to assume that the probability distribution of y_r, given y_{r+1}, is completely defined by its mean and variance.

One must of course distinguish these one-period conditional variances from those of future incomes conditional upon, say, N periods to go. These we can expect to increase through time, though of course this depends on the sequences ϕ_r and σ_r^2. Thus, under our strict-sense Markov assumptions:

$$\mathrm{Var}\,(y_N \mid y_{N+1}) = \sigma_N^2$$

$$\mathrm{Var}\,(y_{N-1} \mid y_{N+1}) = \phi_{N-1}^2\sigma_N^2 + \sigma_{N-1}^2$$

$$\mathrm{Var}\,(y_{N-2} \mid y_{N+1}) = \phi_{N-2}^2\phi_{N-1}^2\sigma_N^2 + \phi_{N-2}^2\sigma_{N-1}^2 + \sigma_{N-2}^2$$

and so on.

We assume that the consumer's intertemporal utility function is additive, of the form:

$$u(C_N, C_{N-1} \ldots C_0) = \sum_{t=0}^{N} \delta^t u_{N-t}(C_{N-t}) \qquad \text{all } N \tag{2}$$

Here δ is a discount factor. No particular supposition is made about the one-period utility functions.

Suppose that the consumer has N periods to go. Then his *state-vector* will comprise his wealth x_N and his last observed income y_{N+1}. We now trace through the solution process to the dynamic programming problem of maximizing the expected value of (2), subject to a given state-vector (x_N, y_{N+1}) and an uncertain income stream $y_N, y_{N-1} \ldots y_1$.

Thus we start off by imagining ourselves in the last period, with state-vector (x_0, y_1). There is now no problem; utility with zero periods to go is fixed at $u_0(x_0) = u_0(r(x_1 - C_1) + y_1)$.

Imagine now one period to go. Then the maximal utility J_1 is given by

$$J_1 = \max_{0 \le C_1 \le x_1} \left[u_1(C_1) + \delta \underset{y_1}{E} u_0(r(x_1 - C_1) + y_1) \mid x_1, y_2\right] \tag{3}$$

The conditional expectation on the right-hand side can be replaced by its certainty equivalent:

$$Eu_0(r(x_1 - C_1) + y_1) = u_0(r(x_1 - C_1) + \phi_1 y_2 - \alpha_1)$$

where α_1 is a one-period risk premium which depends on x_1, y_2, C_1 and σ_1^2. Carrying out the maximization based on (3), we obtain:

$$\alpha_1 = \alpha_1(x_1, y_2, \sigma_1^2) \tag{4a}$$

$$C_1 = C_1(x_1, y_2, \sigma_1^2) \tag{4b}$$

and

$$J_1 = J_1(x_1, y_2, \sigma_1^2) \tag{4c}$$

Now suppose that there are two periods to go. Then

$$J_2 = \max_{0 \le C_2 \le x_2} \left[u_2(C_2) + \delta \underset{y_2}{E} J_1(r(x_2 - C_2) + y_2) \mid x_2, y_3\right]$$

Once again, the expectation can be replaced by a certainty equivalent by subtracting from y_2 a risk premium dependent upon x_2, y_3, C_2, σ^2 and σ_1^2. Solving the resulting (one-period) maximization problem, we obtain

$$\alpha_2 = \alpha_2(x_2, y_3; \sigma_1^2, \sigma_2^2) \tag{5a}$$

$$C_2 = C_2(x_2, y_3; \sigma_1^2, \sigma_2^2) \tag{5b}$$

$$J_2 = J_2(x_2, y_3; \sigma_1^2, \sigma_2^2) \tag{5c}$$

Carry the process back to stage $(N-1)$. We obtain:

$$J_{N-1} = J_{N-1}(x_{N-1}, y_N; \sigma_{N-1}^2, \sigma_{N-2}^2 \ldots \sigma_1^2)$$

The maximization process with N stages to go is now:

$$\begin{aligned} J_N(x_N, y_{N+1}) &= \max_{0 \le C_N \le x_N} (u_N(C_N) + \delta \underset{y_N}{E} J_{N-1} \\ &\quad \cdot (r_N(x_N - C_N) + y_N) \mid x_N, y_{N+1}) \end{aligned} \tag{6}$$

Once again, we can write

$$\underset{y_N}{E}(J_{N-1}) = J_{N-1}(r(x_N - C_N) + \phi_N y_{N+1} - \alpha_N) \tag{7}$$

where after the maximizing process has been carried out,

$$\alpha_N = \alpha_N(x_N, y_{N+1}; \sigma_N^2, \sigma_{N-1}^2 \ldots \sigma_1^2) \tag{8}$$

These equations will serve as the basis for discussion in Section III. For the moment we notice that the above sequence of stages can be taken not only as a computational algorithm, but in reverse order as representing the actual decisions taken over the unfolding of the consumer's lifetime. Suppose for instance, that he has two periods to his horizon. Then his planned consumption in the following period is

$$C_1^{(2)} = C_1(x_2, y_3; \sigma_2^2, \sigma_1^2)$$

This is in general different from C_1 as planned a year later, which is (4b):

$$C_1^{(1)} = C_1(x_1, y_2; \sigma_1^2)$$

The consumer revises his plan as more information becomes available—namely the actual value of y_2. And this is the conclusion that Nagatani is at pains to establish.

II. Risk Premiums

Let us return to equations (6)–(8). We are now N years off our horizon, and we know y_{N+1}. Let C_N^* denote the value of C_N that maximizes (6). Then α_N^* is found from equation (8). The consumer thus behaves as if his coming income is not its expectation $\phi_N y_{N+1}$, but

$$\overset{*}{y}_N = \phi_N y_{N+1} - \overset{*}{\alpha}_N$$

This, together with

$$\overset{*}{x}_{N-1} = r(x_N - \overset{*}{C}_N) + \overset{*}{y}_N$$

is the new state-vector so far as the period $(N-1)$ maximization is concerned.

Now go forward again and consider the equation which (as in (4) and (5)) defines the certainty equivalent for J_{N-2}:

$$(9)\quad \underset{y_{N-1}}{E}\left[J_{N-2}(r(\overset{*}{x}_{N-1} - \overset{*}{C}_{N-1}) + y_{N-1}) \,\middle|\, \overset{*}{x}_{N-1}, \overset{*}{y}_N\right] = J_{N-2}(r(\overset{*}{x}_{N-1} - \overset{*}{C}_{N-1}) + \phi_{N-1}\overset{*}{y}_N - \overset{*}{\alpha}_{N-1})$$

The consumer thus behaves as if his income in period $N-1$ is

$$\begin{aligned}\overset{*}{y}_{N-1} &= \phi_{N-1}\overset{*}{y}_N - \overset{*}{\alpha}_{N-1} \\ &= \phi_{N-1}\phi_N y_{N+1} - \phi_{N-1}\overset{*}{\alpha}_N - \overset{*}{\alpha}_{N-1} \\ &= E(y_{N-1} \mid y_{N+1}) - \phi_{N-1}\overset{*}{\alpha}_N - \overset{*}{\alpha}_{N-1}\end{aligned}$$

The way in which successive certainty equivalents are built up should now be clear. Thus we write

$$\begin{aligned}\overset{*}{y}_{N-2} &= \phi_{N-2}\overset{*}{y}_{N-1} - \overset{*}{\alpha}_{N-2} \\ &= E(y_{N-2} \mid y_{N+1}) - \phi_{N-2}\phi_{N-1}\overset{*}{\alpha}_N \\ &\quad - \phi_{N-2}\overset{*}{\alpha}_{N-1} - \overset{*}{\alpha}_{N-2}\end{aligned}$$

The risk premiums in each period are thus:

$$\begin{aligned}(10)\quad \Pi_N &= \overset{*}{\alpha}_N \\ \Pi_{N-1} &= \phi_{N-1}\overset{*}{\alpha}_N + \overset{*}{\alpha}_{N-1} \\ \Pi_{N-2} &= \phi_{N-2}\phi_{N-1}\overset{*}{\alpha}_N + \phi_{N-2}\overset{*}{\alpha}_{N-1} + \overset{*}{\alpha}_{N-2}\end{aligned}$$

Because the α_r^* are derived from a series of one-dimensional problems, they are unique. So, therefore, are the required premiums Π_r. They obey the recurrence relation

$$(11)\quad \Pi_r = \phi_r\Pi_{r+1} + \overset{*}{\alpha}_r \quad r = (N-1)\ldots 1$$

The simple form of the above relationship masks the fact that the α_r^* are in fact complicated functions; in general α_r^* will depend on α_{r-1}^*, and both will be dependent upon the optimal consumption program. We shall see this more clearly shortly, but in the meantime we should note that the time profile of the risk premiums depends on both the coefficients ϕ_r and the "marginal" premiums α_r^*. It is not necessarily the case that the risk premiums constitute a monotonically increasing function of time.

Now recall that the $J_N, J_{N-1} \ldots J_1$ are all utility functions, defined over the appropriate time interval. They are functions of wealth, evaluated on the supposition that the optimal allocation of future consumptions has been made. It is not unreasonable to assume that these functions exhibit decreasing risk aversion. But it probably is unreasonable to assume that the form of this aversion as a function of initial wealth x is as simple as $1/x$; indeed Pratt showed that the two functions *log* x and $(1/\gamma)x^\gamma$ are the *only* functions exhibiting this simple behavior. So to preserve simplicity in exposition, while hedging on the exact form of decreasing risk aversion, I shall assume that:

From (7),

$$\text{(12a)}\quad \overset{*}{\alpha}_N = \overset{*}{\alpha}_N\left(\overset{2}{\sigma}_N, \frac{1}{r(x_N - C_N^*) + \phi_N y_{N+1}}\right)$$

From (9),

$$\text{(12b)}\quad \overset{*}{\alpha}_{N-1} = \overset{*}{\alpha}_{N-1} \cdot\left(\overset{2}{\sigma}_{N-1}, \frac{1}{r(\overset{*}{x}_{N-1} - \overset{*}{C}_{N-1}) + \phi_{N-1}\overset{*}{y}_N}\right)$$

and so on, each function increasing with respect to both arguments.

We can now see quite clearly that there is nothing exogenous about the α_r^*, and hence the Π_r. They depend very much on the optimal consumption path; or more accurately, they and the consumption path are jointly determined. Thus any calculus that treats them as exogenous is at least incomplete.

In the formulations (12a) and (b), we see that the marginal premiums α_r^* will decrease in size with planned wealth. This seems eminently sensible. Thus in the negative aspect, one would expect someone who consumes most of his income as it comes to hand to be much more sensitive to risk.

Let us imagine what is probably a very common pattern, where wealth is small in the first few years of the life cycle, rises to a peak somewhere in middle age and decreases during retirement years. The marginal premiums α_r^* will therefore exhibit an inverse movement (abstracting from the influence of the σ_r^2). These movements in wealth are probably the result not only of savings plans but also of a roughly coincident movement in income. However, after early middle age has been reached, we will suppose income to be increasing much less quickly. Thus in the years of middle age the wealth effects on the Π_r (via smaller α_r^*) might dampen the income effects (via the ϕ_r). In these circumstances, we might end up with a typical profile for the risk premiums which looks as in Figure 1.

Region *AB* is a high risk region. Rising income (ϕ_r) reinforces large marginal premiums (α_r^*) due to a smaller wealth. Region *CD* is another. Here income expectations are static but on the other hand wealth is now being continuously run down (high α_r^*).

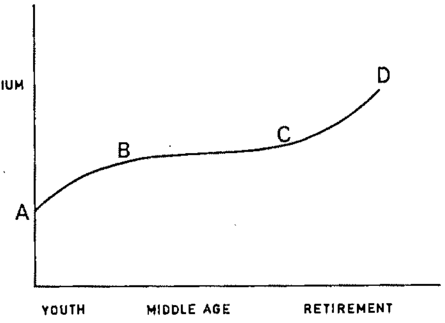


FIGURE 1

It is worth noting at this juncture that an alternative measure of the effect of uncertainty might be to express the risk premium as a proportion of expected income, rather than in absolute terms. Movements of the kind generating Figure 1 could then become quite graphic.

III. Some Implications

Before we proceed to draw out the implications of a curve like Figure 1 for manifest consumer behavior, a couple of preliminary points are in order.

We have seen that the risk premiums were derived by postulating a certain time-shape of wealth. Since this is at least in part determined by the consumption stream, there is an element of circularity in making inferences about the effect on the consumption stream of such a curve! This is just a reflection of the joint determination of consumption plans and risk premiums already referred to. Provided, however, we recognize the existence of a simultaneity problem, the observations below should still prove useful as incorporating one of the directions of causation.

Secondly, the model as outlined in Section II made no allowance for borrowing. If the individual now has the facility for getting into debt, we might expect a reinforcement of any risk aversion tendencies. Nobody likes the embarrassment of having the television set repossessed.

What, then, are the inferences we can draw, on the basis of a curve like Figure 1, about consumption plans under uncertainty as compared with those under certainty;

where the actual income flow in the latter case is the same as the expected flow in the former? To help answer this, we can exploit the dual nature of Figure 1. On the one hand it represents a profile of risk premiums from the vantage point of someone starting out on his life cycle. But we can also imagine the consumer to be moving along the curve with the passage of time; provided, that is, his ideas of the income-generating mechanism (ϕ_r and σ_r^2) remain unchanged. Thus on this latter interpretation, by the time the individual gets to middle age, his risk premiums for the years immediately ahead are (on the appropriate vertical translation of the horizontal axis) small and increasing only slowly.

We can then see that such a curve might produce a double-barrelled Nagatani effect. The individual at point A would consume less than in the certainty case. Thus by the time he reaches middle age he has accumulated more wealth. But middle age is point B, the start of the region of low risk aversion. Both factors reinforce one another and extra consumption results. Until, that is, he reaches C. With declining wealth he becomes more risk averse, and this has a dampening effect on consumption.

Much the same sort of effect may occur with the introduction of premiums in response to uncertainties in rates of return on earning assets. In this case the uncertainties would influence not only the consumption path but the relative amounts held of the various earning assets, and the order with which they are acquired in the course of the life cycle. Thus the individual in his early years might be risk conscious to the extent of foregoing equities altogether in favor of savings deposits. Such holdings would come later with wealth and decreasing risk aversion.

Uncertainties in rates of return to human and nonhuman wealth (and other contingent uncertainties) would in turn have implications for the holding of consumer durables. Such assets are characterized (a) by their indivisibility, which often necessitates debt-financing, and (b) by the fact that, except possibly in a regime of rapidly rising goods prices, there is likely to be a loss involved in any forced realization at a later date (the "locking-in" effect). We should therefore expect both the desired stock of durables and the timing of their acquisition to be markedly sensitive to uncertainty.

In turn, this would have implications for the reaction of the consumer to a windfall. Reverting to the framework of Section II, suppose that with N periods to go, the certainty equivalent income in the coming period is y_N^*. But imagine that the actual income receipt turned out to be y_N. Then his windfall, I define as $y_N - y_N^*$. (Note on this definition that if y_N coincided with $E(y_N \mid y_{J+1})$, there would still be a windfall, of the amount of the risk premium α_N^*.)

A large positive windfall would have two effects. First it would raise his wealth (or his "permanent" income); and secondly it would also, via the increased wealth, bring him on to a less risk-averse position on the risk premium curve. The probable result would be some increase in the desired stock of durables (which would probably be related more to permanent income) but more particularly an advance in the planned timing of their purchase. Windfalls would thus have a special relationship with durables *purchases*. A relationship of this kind has long been suspected, and the present approach to the definition of windfalls may turn out to be a fruitful one on this count.

REFERENCES

K. Nagatani, "Life Cycle Saving: Theory and Fact," *Amer. Econ. Rev.*, June 1972, *62*, 344–53.

E. S. Phelps, "The Accumulation of Risky Capital: A Sequential Utility Analysis," *Econometrica*, Oct. 1962, *30*, 729–43.

J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan.-Apr. 1964, *32*, 122–36.

Private and Social Rates of Return to Education of Academicians: Note

By LORENZO FIGÀ-TALAMANCA*

In a recent article in this *Review* by Duncan Bailey and Charles Schotta, there are two specification errors that, I believe, can invalidate their policy conclusions. The first, common to the majority of papers on human capital, is that rates of return on education cannot be measured by wage differentials. The second is that Bailey and Schotta failed to take into account the peculiar position of universities in the market for academicians. As shown in Table 1 of their paper, the universities enjoy a nearly monopsonistic position in the market for Ph.D. recipients. Therefore a tradeoff will result between high salaries and no subsidies to graduate education on one side, and low salaries and subsidies to graduate education on the other side.

As for the first point, it is by now a commonplace in economic literature, that the social marginal product cannot be measured through wages. For instance Joan Robinson, in her 1971 Richard T. Ely lecture says:

> . . . the *real* wage of each type of labor is believed to measure its marginal product to *society*. The salary of a professor of economics measures his contribution to society and the wage of a garbage collector measures his contribution. Of course, this is a very comforting doctrine for professors of economics, but I fear that once more the argument is circular. There is not any measure of marginal products except the wages themselves. [p. 9]

It is a simple corollary that if wages don't measure marginal product and if "everyone can see that his relative earnings depend on the bargaining power of the group he belongs to" (Robinson), wage differentials cannot be taken as measures for social returns on education.

That the determining factors in wage differentials lay outside education can also be seen, by noticing, that, according to Edward Denison, the average salary of a bachelor's degree holder is 2.35 times the average wage of a worker with just an elementary education, while, according to Robert Solow, the average wage of a worker in the petroleum industry is four times the average wage of a worker in the tobacco industry.

Now, if we accept that the bargaining power of the group is the main factor in explaining wage differentials, professors are, indeed, in a very poor situation. In fact: 1) university professors are, in general, reluctant to unionize; 2) university education is, by and large, a labor intensive activity; 3) a strike and the consequent absence from work will not damage equipment (as in the petroleum industry); 4) the products of the education process are not sold on a piece basis, so that discontinuing production (teaching) for a short time will not financially damage the universities (presumably tuition fees and other income will not go down for a few weeks strike); 5) it is not obvious that consumers (students) buy the direct product of the professors' work (knowledge) and not something (degrees and diplomas) which to some extent can be directly produced by the administration.[1]

---

* Prior to his death in November 1972, Dr. Figà-Talamanca was a member of the Istituto di Economia e Finanza, Facoltà di Giurisprudenza, Università di Roma.

[1] For example, the experience of the high school teachers' union in Italy has shown that strikes have some effect only when they occur during exam and grading periods. To postpone exams annoys the students and consequently the community. An additional holiday, due to teachers' strikes, during normal teaching periods, just cheers the students and consequently the community. I know that in the United States the situation is different for high schools, but it might be similar for colleges and universities. Another example of the relative unimportance of abstension from teaching activities is given by the case of Italian teachers of medical schools. They recently obtained a salary raise and, for the first time, a salary scale higher than that of teachers of other university schools. They obtained

These reasons may explain why the bargaining power of academicians is poor and, as a consequence, why their salaries may appear low, when compared on a human capital basis with the salaries of the other bachelor degree's holders, who, as a rule, operate in a completely different setting, and to whom the reasons listed above do not apply in general. On the other hand, if for some reason a professors' union acquired all at once a strong bargaining power, and as a consequence, professors' salaries increased by 30 percent, then, in the Bailey and Schotta framework, a 6 percent social rate return to investment in graduate education would be secured. But even if we accept the naive neoclassical hypothesis on which, unfortunately, the majority of papers on human capital are based,[2] Bailey and Schotta's analysis of social returns does not hold.

We have in fact a flow of services to society, namely the teaching and research work performed by the Ph.D. recipients, and a monetary flow from taxpayers and consumers (tuition fees). The latter flow can be completely allocated in salaries or partially in salaries and partially as a contribution to human capital formation. Now, if only a small number of Ph.D. recipients went back to work for publicly funded institutions, then the public contribution to graduate education, and hence to human capital formation, would actually represent a subsidy to those private activities that use Ph.D. recipients. But in fact, as shown by Bailey and Schotta, 86 percent of the Ph.D. recipients go to work in publicly funded institutions; that is, the prime employers of Ph.D.s are the producers.[3] Therefore, the public subsidies of graduate education represent a tradeoff for the lower salaries of academicians.

Indeed, in the presence of transaction costs and/or financial market imperfections and/or risk aversion, it might very well be a better allocation of taxpayer's money to subsidize graduate schools and pay lower salaries rather than to pay high salaries to Ph.D. recipients, without subsidizing graduate schools. The first is in fact the policy adopted by other institutions (such as the Armed Forces and their schools) who have a completely monopsonistic position on the output of their training programs.

Therefore, the fact the "rates of return" in arts and education, where 99 percent of the Ph.D. recipients work in publicly funded institutions, are close to zero, may just be a consequence of efficient allocation of taxpayer's money. On the other hand it is natural to conjecture that if Bailey and Schotta had enlarged their inquiry to include those graduate schools (medicine and law) whose output (M.D.s and lawyers) is, by and large, employed in the private sector, they would have found higher rates of return.

One could object to my reasoning by remarking that Ph.D. granting institutions are not the only ones employing Ph.D.s as teachers. Thus graduate schools which produce many more Ph.D.s than they can possibly employ, would not be acting in their own interest, but would subsidize lesser schools who employ Ph.D.s. But first of all, as Bailey and Schotta have shown,[4] the existence of a graduate school does lower the cost of undergraduate education. Second, resource allocation is often decided by an authority which is responsible for institutions of different levels. Moreover, scrapping policies depend on quality standards, therefore major universities must produce a high number of Ph.D.s in order to be able to select among them, four or five years after they completed their degree, the "highly reputed scholars" who are allowed to achieve tenure in a major university. It seems therefore that the Ph.D.s employed in other colleges and universities may be considered to

---

their raise, of course, with a strike, but they did not go on strike as teachers (abstaining from teaching), but rather as physicians of the university hospitals (that is, they refused to treat patients except for emergencies).

[2] For a more detailed critique of the wage differentials approach to human capital growth, see my 1971 paper.

[3] The percentage of Ph.D.s really employed in private activities is probably less than the 14 percent shown by Bailey and Schotta. A large part of natural science Ph.D.s employed in private business is in fact likely to be employed by firms that are only fictitiously private; that is firms whose output is, in total or in part, government research and development projects, paid on a cost plus "fair profit" basis.

[4] See page 22, footnote 4.

some extent a by-product of the production of "highly reputed scholars."[5]

I must recognize that the analysis sketched in this note is far from complete. However I cannot avoid feeling that the naive and crude tools used in the Bailey and Schotta paper cannot lead to any sound policy conclusions. Despite their painstaking work of collecting a very useful set of data, I believe that a more thorough theoretical model should be built before any practical suggestion could be drawn.

REFERENCES

D. Bailey and C. Schotta, "Private and Social Rates of Return to Education of Academicians," *Amer. Econ. Rev.*, Mar. 1972, *62*, 19–32.

E. F. Denison, *The Sources of Economic Growth in the United States and the Alternative before Us*, New York 1962.

L. Figà-Talamanca, "Osservazioni sul valore conoscitivo delle tecniche del "Growth Accounting" ai fini della programmazione economica," *Stato Sociale*, Jan. 1971, pp. 3–27.

J. Robinson, "The Second Crisis of Economic Theory," *Amer. Econ. Rev. Proc.*, May 1972, *62*, 1–9.

R. Solow, "Capital, Labor, and Income in Manufacture" in *The Behavior of Income Shares*, Nat. Bur. Econ. Res. *Stud. Income and Wealth*, New York 1964.

---

[5] As a matter of fact, the same reasoning can be applied to graduate students that are "scrapped" before reaching the Ph.D. (the dropouts), and an analysis of their employment situation has to be carried out before simply stating that they are scrapped and consequently not sold in the market. Or do Bailey and Schotta believe that nothing is added to human capital with one or two years of graduate education when a Ph.D. is not attained?

Private and Social Rates of Return to Education of Academicians: Note

By John A. Tomaske*

In the March 1972 issue of this *Review*, Duncan Bailey and Charles Schotta published a study of the returns to investment in the graduate education of Ph.D. academicians in the United States. They use as a proxy for the academic income of Ph.D. faculty, the salary data reported by the American Association of University Professors (*AAUP*). The alternative income used is an estimate of the income of bachelor's degree holders taken from an extensive survey of salaries in occupations open to bachelor's degree holders in the state of California. They compute both a private and a social rate of return. Their private rate of return includes as costs an estimate of the foregone income of graduate students. Their social rate of return in addition to income foregone by students, includes an estimate of the per unit contribution of the state of California to graduate education at the Berkeley and Los Angeles campuses of the University of California. A second social rate of return is computed by including an estimate of the costs of graduate school dropouts. Each of these rates of return is estimated for two through six year periods of time spent in graduate school and for eleven different academic income patterns.[1]

Bailey and Schotta explicitly omit from their calculations all incomes earned by academicians in addition to their academic year contract salaries. Since they are interested only in test-of-the-marketplace conclusions, they do not consider externalities or the public good aspects of research. Their conclusions may be summarized as:

1. The private and social rates of return are, in general, either zero or less than 1 percent. The private incentive to invest in graduate education is explained by psychic income and the attractiveness of academic life styles.
2. The findings of other studies that the return to undergraduate education is in the neighborhood of 10 percent implies, to the authors, that resource allocation would be improved by reallocation of resources away from graduate to undergraduate education. They hazard a "guess" that the resources removed from graduate education should be in the humanities, education, and certain of the social sciences.
3. The authors believe that an important area for cost saving is reduction of both the graduate school dropout rate and the time spent in graduate school to obtain the Ph.D.

The purpose of this note is to argue that Bailey and Schotta's approach and the data they employ understates the return to the graduate education of the Ph.D. academicians and that their conclusions are seriously misleading as a guide to public policy.

I

The *AAUP* is concerned with the economic condition of all faculty employed by institutions of higher education in the United States. *AAUP* Committee Z does not distinguish between the salary status of Ph.D. and non-Ph.D. faculty when reporting average earnings. The salary data used by Bailey and Schotta as a proxy for the academic income of Ph.D. faculty are an average by academic rank of the salaries paid to full-time academic personnel regardless of degree level or institutional affiliation. Of the 936 institutions included in the *AAUP* self-grading compensation survey for 1966 used by Bailey and Schotta, 11.9 percent were universities, 63.2 percent liberal arts colleges, 5.2 percent teachers colleges, 14.8 percent junior colleges, 2.8 percent technical institutes, and 1.9 percent fit none of these categories. Of these institutions, 862 institutions employing 181,467 full-time faculty could be classified by the average compensation grades used by *AAUP* at that time. Institutions with an average grade of D or below employed 21.9 percent of the full-time faculty, C graded institutions employed 37.4 percent, and institutions graded A or B employed 40.6 percent of the full-time faculty. One institution employing 174 faculty was given an average grade of AA.

---

* Associate dean of academic planning (Graduate Studies and Research), California State University, Los Angeles. I have benefited from suggestions by Judd Hammack, George Jensen, Donald Moore, and Richard Roseman, and by an unnamed referee on an earlier draft of this note.

[1] The basic salary stream for academically employed doctorates was determined by using the University of California salary scale and promotion schedule as a bench mark. This reference income stream was converted to ten income streams using *AAUP* salary data and *AAUP* ratings by academic rank. An eleventh salary pattern is the weighted average of the average annual incomes of patterns one through ten. The weights were the percentage of faculty in 829 institutions which occupied that salary pattern.

Unfortunately, there are no data available concerning the income of Ph.D. academicians which are comparable in detail to the salary information published by *AAUP*. However, a study by Ralph Dunham, Patricia Wright, and Marjorie Chandler, published by the U.S. Office of Education and based on a survey conducted spring 1963 covering universities and four-year colleges employing 138,200 full-time faculty can be used to indicate the direction and gain some idea of the magnitude of the bias contained in the salary data used by Bailey and Schotta.

Only 51 percent of the full-time faculty employed by universities and four-year colleges held the Ph.D. These individuals were concentrated in the associate and full professor ranks and in larger institutions. The influence of non-Ph.D. faculty salaries on reported salary averages appears greatest in the instructor and assistant professor ranks. These data do not distinguish between Ph.D. and non-Ph.D. faculty by salary levels. The extent of the premium paid to doctorates in four-year institutions with substantial numbers of non-Ph.D. faculty is not known. Obviously, the salary averages reported by these institutions will be increasingly less representative of Ph.D. salaries as the proportion of nondoctorates in the faculty increases. *AAUP* salary data are a highly imperfect proxy for the income of doctorate faculty. The use of these data significantly understates the academic year contract salaries paid to doctorates employed in institutions of higher education.

TABLE 1—TEACHING FACULTY WITH THE DOCTORATE: SELECTED CHARACTERISTICS, UNIVERSITIES, AND FOUR-YEAR COLLEGES, SPRING 1963

	Number	Percent of Total Faculty
Total Doctorates	69,949	51
Rank		
Professor	29,108	78
Associate Professor	20,105	61
Assistant Professor	16,908	42
Instructor	2,159	10
Other	1,669	33
Selected Institutional Attributes:		
Universities	40,525	59
Public	26,224	58
Private	14,301	60
750 Faculty or More	18,896	65
Under 750 Faculty	21,629	54
Colleges and Technological Institutions:	29,423	42
Public	13,611	41
Private	15,812	44
200 Faculty or More	8,596	51
Under 200 Faculty	20,827	40

Source: Dunham, Wright, and Chandler, Table 1.

Bailey and Schotta explicitly omit income earned by faculty from summer teaching, royalties, consulting, and other professional activities not included in reported academic year contract salaries. N. Arnold Tolles and Emmanuel Melichar have analyzed income supplement data for economists collected by *The National Register of Scientific and Technical Personnel*. They found that 83.2 percent of economists employed by educational institutions reported earned income supplements. Those economists employed on a nine- to ten-month academic year basis had supplemental incomes equal to 29.4 percent of their basic salary in 1964. Dunham, Wright, and Chandler found in 1963 that 81 percent of all doctorate faculty in all disciplines on academic year contracts

TABLE 2—ACADEMIC YEAR CONTRACT SALARY OF TEACHING FACULTY 1962–1963: SELECTED CHARACTERISTICS[a]
(84,784 Faculty)

	Doctorates			Nondoctorates		
	Percent	Mean Income	Median Income	Percent	Mean Income	Median Income
All Ranks	49.3	9,490	9,099	50.7	7,146	6,913
Professor	78.2	11,299	11,078	21.8	10,901	10,621
Associate Professor	61.3	9,053	9,058	38.7	8,072	8,093
Assistant Professor	40.9	7,669	7,526	59.1	6,895	6,948
Instructor	9.2	6,675	6,851	90.8	5,644	5,753
Other	31.6	7,848	7,437	68.4	6,838	6,667

[a] Includes faculty on 9–10 month academic year contracts only.
Source: Dunham, Wright, and Chandler, Table 38.

earned income supplements equal to 25.3 percent of their contract salaries. Nondoctorate faculty were less able to augment their academic year contract salaries. Only 66 percent earned outside income. Supplemental income was earned by a majority of faculty in all discipline categories except home economics and by faculty in all academic ranks.

II

In estimating a social rate of return to the graduate education of academicians, Bailey and Schotta use, without discussion, state of California budget data and University of California budget formula to obtain a per unit measure of the state contribution. Discussion of the conceptual and operational difficulties involved in segregating the cost of graduate education in universities which jointly produce graduate education and research as well as undergraduate instruction is outside the scope of this note and may be found elsewhere.[2] The California budgeting process, in common with that found in other states, involves much that is traditional and more that is negotiated. John H. Powell and Robert D. Lamson, in a study sponsored by the Council of Graduate Schools in the United States and the National Association of College and University Business Officers, found that differences in budgeting processes and accounting methodology precluded institutional comparisons of the costs of graduate education in any but the most imprecise terms. It is not now possible to determine how accurate or representative Bailey and Schotta's cost estimates may be.

Bailey and Schotta investigate the effects of graduate school dropouts on their computed social rates of return. They state, "If, as is well known in industrial quality and cost control, some defects are produced which are not subsequently sold and must be scrapped, then the resources expended on their production must be averaged over acceptable units, thus increasing their average costs (p. 27)." Their dropout hypothesis implies a 50 percent rejection rate spread over a four-year graduate education. It is assumed that these dropouts do not benefit economically from exposure to graduate education. This assumption is either contrary to what is generally believed to be the nature of the educational transformation process or assumes that certification rather than education is of long-run economic importance.

In any case, we do not know what happens to the graduate school dropout and there is very little information on the economic condition of the terminal master's degree holder.[3] However, Dunham, Wright, and Chandler find that 32 percent of all teaching faculty at universities and four-year colleges either have a terminal master's degree or have taken work beyond the master's degree but have not obtained the Ph.D. Bailey and

---

[2] See, for example, W. Lee Hansen and Burton Weisbrod, Marc Nerlove, and John Powell and Robert Lamson.

[3] Some information is provided by Orley Ashenfelter and J. D. Mooney, and Lynn Maxwell.

TABLE 3—OUTSIDE EARNINGS OF ACADEMIC YEAR CONTRACT TEACHING FACULTY: 1961–1962

	Percent of Academic Year Faculty with Outside Earnings	Mean Earnings	Median Earnings
Education Level			
Doctorates	81	2,399	1,810
Non Doctorates	66	1,857	1,252
Rank			
Professor	80	3,026	2,036
Associate Professor	78	2,269	1,731
Assistant Professor	74	1,700	1,359
Institutional Affiliation			
Universities	79	2,577	1,847
Colleges	69	1,790	1,321

Source: Dunham, Wright, and Chandler, Table 40.

Schotta have computed social rates of return by adjusting their cost data for graduate school dropouts but without adjusting their income data to a similar basis.

Bailey and Schotta mention the word research once and ignore it completely in their concluding statements. This is unfortunate since research is both a part of graduate training and a portion of the output of the academically employed Ph.D. recipient.[4] Apparently, the social value of the research output of faculty is assumed to be reflected in the salaries paid academicians. The information which results from academic research is, with few exceptions, a public good. The value of public goods is generally considered as not measured by market prices.[5] Bailey and Schotta offer no justification for their assumption that the academic marketplace is an exception.[6] This is of particular importance in the context of graduate education since a major rationale for subsidizing this activity has historically been the additional knowledge generated by these highly trained individuals.

There is little documented evidence concerning the magnitude of the benefits of academic research. Zvi Griliches has estimated a social rate of return of 700 percent on the resources invested in the development of hybrid corn. His estimate of the net social benefits derived from this single commodity is in excess of $240 million annually after 1955.[7] Burton Weisbrod estimated a most probable social rate of return of 11 to 12 percent on research costs resulting in the development of poliomyelitis vaccine. Although both of these developments involved substantial academic research and resulted in a salable commodity, there is no evidence that a significant portion of these benefits was captured by either individual scholars or academic institutions.[8]

III

Bailey and Schotta estimate the private rate of return to investment in graduate education of doctorate faculty is either zero or less than 1 percent. Aside from the shortcomings of their estimates of the academic year salaries of Ph.D.s, the omission of supplemental income substantially understates the annual income of academicians. Using their weighted average salary pattern eleven and assuming four years spent in graduate school, they show that academicians must earn supplemental income equal to approximately 15 percent of their academic year salaries each year to attain a private return of 7 percent. A 10 percent return requires supplemental income equal to 23.5 percent of academic year salaries. Since supplemental incomes in this range are common, it is unnecessary to use psychic income or life style effects to explain the incentive for individuals to invest in graduate education in 1966.

---

[4] The published output of teaching faculty with the doctorate is substantial. Dunham, Wright, and Chandler found that 81 percent reported publishing one or more articles and 43 percent reported five articles or more. Thirty-four percent reported they had published professional books.

[5] The issue is more complex. Kenneth Arrow shows that basic research which is primarily an input for further research is particularly unlikely to be rewarded.

[6] Those readers who have participated in hiring, promotion, retention, and tenure decisions may wish to reflect on the efficiency of this "market."

[7] Griliches estimates an annual rate of 35 to 171 percent per dollar spent on all agricultural research. He estimates that one-half of these research expenditures were from public sources.

[8] The tomato harvester and biologically suitable tomato plants were developed primarily through a joint effort by the University of California and private industry. Andrew Schmitz and David Seckler note that the University of California owns some patent rights and has captured a small portion of the substantial social benefits of this research.

The social rate of return presented by Bailey and Schotta should be viewed skeptically. Aside from the aforementioned understatement of the academic incomes of doctorates, their concept of social benefits is too restricted. If only those social benefits priced in a market are included, graduate education may appear to be a less efficient use of resources merely because a larger proportion of the benefits of undergraduate instruction or other public investment options can be measured by market prices. Estimates of both the social costs and benefits of graduate education must be improved before policy pronouncements are appropriate.

The authors "guess" that overinvestment has occurred in the humanities and "soft" sciences. Before this intuitive assertion can be accepted, knowledge of both the costs and the nature of the social and private benefits of these disciplines is needed. Powell and Lamson found evidence of relatively lower instructional costs in these disciplines. The benefits of humanistic knowledge cannot be cavalierly dismissed as consumption goods if only in view of the long-standing academic tradition that this knowledge is necessary both for effective judgmental behavior and the broader perspective which separates the creative scientist from the technician.

REFERENCES

O. Ashenfelter and J. D. Mooney, "Some Evidence on the Private Returns to Graduate Education," *Southern Econ. J.*, Jan. 1969, *35*, 247–56.

D. Bailey and C. Schotta, "Private and Social Rates of Return to Education of Academicians," *Amer. Econ. Rev.*, Mar. 1972, *62* 19–31.

R. Dunham, P. Wright, and M. Chandler, *Teaching Faculty in Universities and Four Year Colleges: Spring 1963*, Washington 1966.

Z. Griliches, "Research Costs and Social Returns: Hybrid Corn and Related Innovations," *J. Polit. Econ.*, Oct. 1958, *66*, 419–31

W. L. Hansen and B. A. Weisbrod, *Benfits, Costs, and Finance of Public Higher Education*, Chicago 1969.

L. Maxwell, "Some Evidence on Negative Returns to Graduate Education," *Western Econ. J.*, June 1970, *8*, 186–89.

M. Nerlove, "On Tuition and the Costs of Higher Education: Prolegomena to a Conceptual Framework," *J. Polit. Econ.*, June 1972, *80*, Part II, S178–S218.

J. H. Powell and R. D. Lamson, *Elements Related to the Determination of Costs and Benefits of Graduate Education*, Washington 1972.

A. Schmitz and D. Seckler, "Mechanized Agriculture and Social Welfare: The Case of the Tomato Harvester," *Amer. J. Agri. Econ.*, Nov. 1970, *52*, 569–77.

N. A. Tolles and E. Melichar, "Studies of the Structure of Economists' Salaries and Income," *Amer. Econ. Rev.*, Dec. 1968, supp., *58*, 1–54.

B. A. Weisbrod, "Costs and Benefits of Medical Research: A Case Study of Poliomyelitis," *J. Polit. Econ.*, May-June 1971, *79*, 527–44.

"Further Progress: The Economic Status of the Profession," *AAUP* Bull., June 1967, *53*, 136–95.

Housing Market Discrimination, Homeownership, and Savings Behavior: Comment

By JOHN F. MCDONALD*

In their recent paper in this *Review*, John F. Kain and John M. Quigley (KQ) state:

> "We are forced to conclude that 'supply restrictions' on Negro residential choice and on the kinds of housing available to black households may be largely responsible for the wide discrepancy between ownership rates for otherwise identical black and white households." [p. 270]

The test which KQ employ to demonstrate the existence of supply restrictions is based on the relative lack of single family houses available to blacks, pp. 270–72. They find, in a sample which uses cities as observations, that the extent to which the actual black homeownership rate falls short of the expected rate is negatively correlated with the proportion of central city housing which is single family.[1] As KQ acknowledge, theirs is not a complete test of the supply restriction hypothesis. This comment provides further tests of that hypothesis.

A relative lack of single family houses in the black housing submarket[2] causes lower black homeownership because apartments are not generally available for owner occupancy and single family houses are usually owner occupied. However, some single family houses in black areas may not be available for owner occupancy due to difficulties which blacks face in the mortgage market. The procedure followed in this paper is to make a correction for housing stock characteristics. Assuming the correction is made properly, the remaining difference between the black and white homeownership rates is due to the fact that blacks have more difficulty obtaining mortgages on single family houses.

Section I presents a linear probability-of-homeownership function with a correction for housing stock characteristics. Section II presents a model in which the type of tenure (own or rent) and the structure type of a household are jointly dependent variables. Section III examines some implications for policy.

I. Probability of Homeownership: Single Equation Model

Column 1 in Table 1 shows a probability-of-homeownership function estimated by ordinary least squares.[3] This equation is simi-

---

* Assistant professor of economics, University of Illinois at Chicago Circle. The work described herein was carried on simultaneously with and independently of the work of Kain and Quigley as part of my doctoral dissertation. I wish to thank Peter Mieszkowski, Robert Goldfarb, Guy Orcutt, and the referee for helpful comments. Views expressed are solely my own.

[1] However, in this test there is no clear reason for including the variable "actual white ownership rate" in an equation which determines the quantity "expected black ownership rate minus actual black ownership rate." KQ find that a larger white ownership rate is associated with a larger difference between actual and expected black ownership rates. KQ argue for including the white ownership rate to control for "intermetropolitan variation in the relative cost of owner-occupied and renter-occupied housing or differences in the timing of urban development" (p. 271). Their statement does not indicate whether the coefficient in question should be positive or negative.

[2] A model which could be thought to underlie the KQ study is a housing market segmented by race.

[3] The data is a stratified sample of 500 Detroit households taken from the Detroit Transportation and Land Use Survey (*TALUS*) of 1965. KQ use household data from St. Louis. Ordinary least squares results are shown to facilitate comparison with the two-stage least squares results in Section II. A generalized least squares procedure yields more efficient estimates. (See KQ, p. 265, and Arthur Goldberger, p. 250.) The weight used for each observation is $(1/P(1-P))^{1/2}$, where P is the probability of homeownership estimated by ordinary least squares. If P is less than zero or greater than one, a weight must be arbitrarily assigned. One procedure is to assign .01 for $P<0$ and .99 for $P>0$. This procedure is used in my case because income and age of household head enter as continuous variables and cause P to exceed 1.0 for some households. (KQ enter income and age is continuous variables, also.) The *GLS* estimate of the coefficient of the race dummy is $-.064$ $(t=1.88)$ instead of $-.100$. The other *GLS* coefficients are very close to those found by *OLS*.

TABLE 1—PROBABILITY-OF-HOMEOWNERSHIP FUNCTION

	(1) *OLS*	(2) *OLS*	(3) *2SLS*
Black household (dummy variable)	−.100	−.064	−.045
	(2.592)	(1.817)	(1.075)
Income (thousands of dollars)	.014	.009	.008
	(3.030)	(2.252)	(1.550)
Length of tenure (years)	.037	.028	.025
	(9.581)	(7.651)	(3.441)
Age of household head	.007	.005	.005
	(5.524)	(3.924)	(1.840)
Years of education, household head	−.019	−.017	−.013
	(4.209)	(3.979)	(2.246)
Size of household (number of people)	.061	.041	.032
	(4.323)	(3.154)	(1.379)
Life cycle variables (dummies)			
Married, no children	−.054	−.038	
	(.951)	(.732)	
Married, youngest child aged 0 to 5	−.030	−.056	
	(.383)	(.792)	
Married, youngest child aged 5 to 17	.010	−.011	
	(.138)	(.166)	
Married, youngest child over age 17	−.111	−.056	
	(1.399)	(.772)	
Structure type variables			
Structure type (1 for single family house or duplex, 0 otherwise)		.376	
		(10.180)	
Two-stage estimate of the effect on homeownership probability of occupying a single-family house or duplex			.503
			(1.824)
Intercept	−.072	−.039	−.091
R^2	.399	.504	
sample size	500	500	500

T-statistics are in parentheses. The probability of homeownership for an unmarried individual living alone is included in the intercept term.

lar to equation (1) in KQ. The dependent variable is dichotomous; zero for renters and one for owner-occupants. The regression shown here in column 1 differs from the KQ model in two respects. One independent variable has been included which KQ do not use; length of tenure in the housing unit. KQ, p. 274, and others state that the relative cost of owning and renting depends crucially on the expected length of tenure; hence actual tenure at the time of the survey has been added as a proxy variable for expected tenure. The sample used here excludes households headed by divorced or widowed women and retired persons on the grounds that current income for such people misrepresents permanent income.

The results in column 1 are much like the KQ results. It is found that blacks are 10 percent less likely than whites to be homeowners after correcting for household characteristics. (KQ found the difference to be 8.8 percent.) The one major difference between the two sets of results is that the size of household variable has a negative coefficient in KQ's results and a positive coefficient in the results reported here.[4] The positive coefficient is more plausible because larger households need more housing space and are more likely to occupy single family houses.

As a first additional test of the supply restriction hypothesis, a structure-type variable has been added to the linear probability function. The variable is dichotomous; one for single family house or duplex and zero otherwise.[5] Column 2 in Table 1 contains the results. It has been assumed that structure type is predetermined in the probability-of-ownership function; that households decide structure type prior to the decision to own or rent. For example, the market first assigns a household to single family houses. The probability of homeownership is partly determined by the fact that the household is restricted to the single family houses. If the 10 percent racial difference in homeownership rates is due only to physical supply characteristics, then the inclusion of the structure-type variable would drive the coefficient of the race dummy to zero. The coefficient of the structure-type variable is highly significant, and the coefficient of the race dummy is reduced from $-.10$ to $-.064$. The calculated t-value for the race dummy is now 1.82, indicating that the coefficient has about a 96 percent probability of being less than zero. Also, the size of household variable continues to make a contribution, an unexpected result if we presume that the influence of household size operates only through the choice of structure type.

II. Probability of Homeownership: Two-Equation Model

Rather than assume that structure type is predetermined in the probability-of-homeownership function, it is more reasonable to assume that structure type and type of tenure are simultaneously determined. The model allows for the possibility that a household which occupies a single family house is more likely to own and a household which owns is more likely to occupy a single family house. The model has two equations because the correlation between structure type and type of tenure is not perfect. In fact, this correlation is hypothesized to be lower for black households than for white households because of difficulties which blacks face in the mortgage market. On the other hand, if the lack of black homeownership is due entirely to a lack of single family houses in the black submarket, then the effect of race would operate only through the probability of occupying a single family house, i.e., the coefficient of the race dummy in the probability of homeownership function would be zero.

If structure type and type of tenure are jointly determined variables, then the ordinary least squares procedure used to compute column 2 of Table 1 yields biased estimates of the coefficients. Such a model has both simultaneity and heteroskedasticity problems.[6] Assuming the distributions of the two error terms are similar to the distribution ordinarily assumed for the single linear probability function, two-stage least squares will provide unbiased and inefficient estimates of the coefficients of the probability of homeownership function.

Consider the following two-equation model.

$$T = a_{12}S + cX_1 + u_1 \tag{1}$$

$$S = a_{21}T + dX_2 + u_2 \tag{2}$$

Here T is type of tenure ($0=$rent, $1=$own) and S is structure type (1 for single family house or duplex, 0 otherwise). The other notation is as follows.

$c, d =$ vectors of coefficients

$X_1, X_2 =$ vectors of predetermined variables

$u_1, u_2 =$ error terms

The error terms are assumed to be distributed as shown in Table 2.

---

[4] When the size of household is removed from the regression, variables indicating the presence of children under age 17 enter with significant positive coefficients.

[5] Duplexes have been classified as being more like single family houses than multiple-unit structures. Duplexes constitute 2 percent of the sample used here.

[6] See Goldberger, pp. 248–50, for a discussion of the heteroskedasticity problem in linear probability functions.

TABLE 2—DISTRIBUTION OF ERROR TERMS

T	S	u_1	$Pr(u_1)$	u_2	$Pr(u_2)$
0	1	$-a_{12} - cX_1$	$\frac{1 - a_{12} - cX_1}{2}$	$1 - dX_2$	$\frac{dX_2}{2}$
0	0	$-cX_1$	$\frac{1 - cX_1}{2}$	$-dX_2$	$\frac{1 - dX_2}{2}$
1	1	$1 - a_{12} - cX_1$	$\frac{a_{12} + cX_1}{2}$	$1 - a_{21} - dX_2$	$\frac{a_{21} + dX_2}{2}$
1	0	$1 - cX_1$	$\frac{cX_1}{2}$	$-a_{21} - dX_2$	$\frac{1 - a_{21} - dX_2}{2}$

Note that $E(u_1) = E(u_2) = 0$ and that $Pr(u_1)$ and $Pr(u_2)$ always fall between zero and one. Also, u_1 and u_2 are not independent. $\text{Cov}(u_1, u_2)$ can be computed by adding together the products of the terms appearing in the last four columns of each row in Table 2: $\text{Cov}(u_1, u_2) = 0$ if $a_{12} = 0$ or $a_{21} = 0$.

The two-stage least squares results are shown in column 3 of Table 1. All the independent variables, including the race dummy, are used to estimate the probability of occupying a single family house in the first stage. In the second stage, the life cycle dummies are excluded from the probability of homeownership function to satisfy the order conditions for identification. In these results, all the coefficients are closer to zero than in column 1. The change in the coefficient of the race dummy from −.10 to −.045 (t = 1.075) means that the effect of being black, after correcting for the effect of housing stock characteristics, is to reduce the probability of homeownership by 4.5 percent. The effect of housing stock characteristics is to reduce black homeownership probability by 5.5 percent. The supply restriction due to a lack of single family houses has an impact on black homeownership, but this factor fails to account for all the racial difference in ownership probability. After correcting for household characteristics, the racial difference in homeownership probability is 10 percent. This difference is reduced to 6.4 percent or 4.5 percent when account is taken of the nature of the housing stock in the black and white submarkets. This remaining difference is associated with the operation of the market for mortgages on single family houses.

The sample used in the regressions was tabulated to discover how many households rent single family houses. The results are shown in Table 3.

TABLE 3—OCCUPANTS OF SINGLE FAMILY HOUSES WHO RENT

Blacks, income under \$5000	40% rent
Blacks, income \$5000 to \$10,000	29.2% rent
Whites, income under \$5000	8.7% rent
Whites, income \$5000 to \$10,000	11.5% rent

III. Summary and Implications for Policy

In a perfectly functioning housing market, we could assume that each household selects the type of tenure which maximizes its utility. People who move frequently, are highly risk averse, or dislike the responsibilities of homeownership, would be observed to be renters. People who desire to consume small amounts of housing services rent because small housing units are not generally available for purchase. If we presume that black and white people do not differ in risk aversion and taste for the responsibilities of homeownership, KQ and this study show that there are some urban black households who rent but would own if they faced the same housing market that whites face. The white housing submarket is characterized by a larger relative amount of single family houses and less difficulty in obtaining mortgage credit. Both factors are important causes of the racial difference in homeownership rates, but KQ provide a statistical test which demonstrates only the effect of the relative lack of single family houses in the black submarket. Part of the lack of black ownership in Detroit in 1965 was due to an inability to obtain mortgages on some of the single family houses which blacks were occupying.

The results of this study suggest that, if the goal of policy is to increase the proportion of blacks who own homes, some increase could be accomplished without moving blacks from their current neighborhoods by making mortgages more readily available for older housing. However, elimination of the statistically corrected racial difference in ownership probability will require elimination of restrictions both in the mortgage market and in the type of housing stock made available to black consumers.[7]

REFERENCES

A. S. Goldberger, *Econometric Theory*, New York 1964.

J. F. Kain and J. M. Quigley, "Housing Market Discrimination, Homeownership, and Savings Behavior," *Amer. Econ. Rev.*, June 1972, *62*, 263–77.

---

[7] Section 235 of the 1968 Housing Act provides below-market interest rates and FHA-guaranteed mortgages for new and renovated houses purchased by people in the $3000 to $7000 income range. This program has been operating in the black housing market in Detroit, and recently a scandal has been uncovered in the Detroit program. According to newspaper accounts, in some cases renovation has meant only that defects in older houses have been covered up. The result has been that new owners have defaulted on their mortgages because of an inability to meet mortgage payments and repair costs. It seems clear that a program which grants mortgages for older houses must provide careful inspection services prior to the sale. However, it may not be possible to administer properly a program of guaranteeing mortgages for older housing. If this is the case, then black homeownership should be increased by opening to blacks more areas of newer single family housing and providing mortgage assistance where needed.

Housing Market Discrimination, Homeownership, and Savings Behavior: Reply

By JOHN F. KAIN AND JOHN M. QUIGLEY*

John McDonald's comment is a useful verification and extension of our earlier analysis of the effects of racial discrimination in urban housing markets. In our paper we demonstrated that black households have a lower probability of home ownership or purchase than otherwise comparable white households. We concluded that this difference results largely from "supply restrictions," i.e., a limited supply of units suitable for home ownership within the ghetto and substantial impediments to the purchase of homes outside the ghetto by black households.

Our paper contained two empirical investigations: a detailed analysis of the probability of home ownership and purchase, based on a large sample of St. Louis households, and an analysis of intermetropolitan differences in black and white home ownership among 18 large metropolitan areas. The second of these two analyses was devised primarily to test the supply restriction hypothesis. In it the proportion of central city units that are single family and the proportion of the *SMSA* black population that reside in the suburban ring were used to represent these supply restrictions.

In his note, McDonald replicates our St. Louis analysis of homeownership using Detroit data and correctly points out that part of the difference he obtains in black and white ownership probabilities may result from the "fact that blacks have more difficulty obtaining mortgages on single family houses." He then attempts to decompose the racial coefficient in our model into two components: 1) a component due to difficulties in obtaining morgages, and 2) a component due to our notion of supply restrictions. In his first alternate model, he includes whether the household lives in a single family unit as an explanatory variable (Equation (2), Table 1). In his second alternate model, structure type and tenure are jointly determined (Equation (3), Table 1). He interprets the racial coefficient in these two alternate probability-of-ownership equations as reflecting problems of access to mortgages.

McDonald does not discuss the nature of the difficulty blacks encounter in obtaining mortgages, but at least two sources come to mind. Blacks may have difficulty obtaining mortgages because of overt racial discrimination by lenders or because they, as individuals, are considered poorer credit risks than comparable whites. Alternatively, mortgages may be unavailable because the dwelling units available to potential black home buyers are old, in poor condition, located in declining neighborhoods, or have other characteristics that make them high risk investments.

Differences in mortgage availability arising from the characteristics of dwellings and neighborhoods are clearly an integral part of our concept of supply restrictions, but overt racial discrimination and differences in the evaluation of credit worthiness by lenders are conceptually distinct.

Unfortunately, McDonald's method fails to distinguish between these two kinds of effects. Indeed, in this respect it goes no further than our analysis of the difference in ownership rates among metropolitan areas. Our analysis of intermetropolitan differences in black and white home ownership rates can be interpreted in terms virtually identical to McDonald's analysis. If the coefficient estimates reported in our equations (4) and (5) are accepted at face value, the results imply that if central city ghettos were composed entirely of single family housing (i.e., if S_c in equations (4) and (5) had a value of 1.0) the mean difference between actual and "expected" black home ownership rates would be reduced to about 3 percentage points

---

* Harvard University.

(using equations (4)) or 6 percentage points (using equation (5)).

For the Detroit metropolitan area, analyzed by McDonald, our estimate of the difference between expected and actual black home ownership rates is 26 percentage points. The regression coefficients from equation (4) indicate that, if Detroit's black ghetto were composed entirely of single family housing this difference would be reduced to 12 percentage points. Finally, the use of the regression coefficients from equation (5) reduces the difference between expected and actual black home ownership rates in Detroit to 6 percentage points. Of course, the assumed change in the characteristics of Detroit's housing stock would imply higher white homeownership rates as well. Thus McDonald's results, using individual households in Detroit as units of observation, are a consistent extension of our own intercity analysis.

We differ more sharply with McDonald's conclusions regarding appropriate public policies. It is hard to disagree with the view that steps should be taken to ensure that banks and other lenders do not discriminate against blacks or other ethnic groups in mortgage dealings. But if the lower ownership rates among blacks living in single family units reflect the higher risks of lending in old, low quality transitional neighborhoods, we are skeptical of programs that would provide mortgages to blacks to buy units in ghetto and boundary areas. Experience with the Boston Banks Urban Renewal Group program in Boston and with similar programs in other cities suggests that such measures in a segregated housing market may operate to consolidate the ghetto.

The obvious public policy alternative is to design and emphasize programs which would make the entire nonghetto housing market more accessible to prospective black home owners. (See Kain and Joseph Persky and our unpublished 1972 paper.) Even if the objective of public policy is no more broad than "to increase the proportion of blacks who own homes" (McDonald, p. 229), the latter policy should provide a much greater payoff.

REFERENCES

J. F. Kain and J. J. Persky, "Alternatives to the Gilded Ghetto," *Publ. Interest,* Winter 1969, *14,* 74–87.

——— and J. M. Quigley, "Housing Market Discrimination, Homeownership, and Savings Behavior," *Amer. Econ. Rev.,* June 1972, *62,* 263–77.

——— and ———, *Discrimination and a Heterogeneous Housing Stock: An Economic Analysis,* unpublished manuscript, 1972.

J. F. McDonald, "Housing Market Discrimination, Homeownership, and Savings Behavior: Comment," *Amer. Econ. Rev.,* Mar. 1974, *64,* 225–29.

The Administered-Price Thesis Denied: Note

By J. Fred Weston, Steven Lustgarten, and Nanci Grottke*

In the opening paragraph of his recent article in this *Review*, Gardner Means makes some sweeping generalizations with regard to the pattern of price behavior during the 1970 recession (1972, p. 292). However, his data cover only the period May 1957 through March 1962. Using Means' earlier criterion for classifying industries, we find that the data for the recent recession as well as for the earlier periods contradict the administered-price thesis.[1]

The empirical tests presented by Means utilized 63 commodities selected by George Stigler and James Kindahl (pp. 22–28) because they were alleged to exhibit administered price behavior. From the 63, Means separates 13 "market-dominated prices" which he defines by three criteria: 1) the 4-firm concentration ratio is low, 2) the market is worldwide, and 3) the industry is raw material dominated. Application of these criteria to the remaining 50 would classify many additional as "market-dominated" commodities so that his results are dependent on the subjective judgments made. Means himself had previously emphasized that concentration alone produced administered-price behavior: "But concentration in industry has brought a new type of inflation . . . " (1964, p. 22).[2] Accordingly, we utilize 4-firm concentration ratios to place the 63 industries into three groupings suggested by the tests set forth in the "Merger Guidelines" of the Department of Justice, pp. 1–27. In Table 1 the truncated version of the administered-price thesis is examined and in Table 2 the full version is tested.[3]

Table 1 shows the distribution of price changes for different levels of seller concentration in each of the four time periods considered by Means. Each panel contains two classifications of price changes which follow from the truncated administered-price thesis prediction of no change or increase during contractions, and no change or decrease during expansion. Chi-square statistics were computed for each of the periods with the null hypothesis that the distribution is the same for all levels of concentration. For the first contraction the differences between the groups are significant at the 20 percent confidence level. For the other three periods the administered-price thesis falls far short of support by the significance tests. Admittedly, the chi-square tests are not statistically powerful, tending to favor the null hypothesis that the administered-price hypothesis is wrong. However, the expected value of chi-square with two degrees of freedom is 2.0, and three of the five individual period results are actually below the expected value so that the weakness of the test is irrelevant. Further, combining the five periods into a single test also results in a chi-

---

* Research Program in Competition and Business Policy, University of California, Los Angeles.

[1] A similar conclusion based on empirical studies was reached by Louis Phlips for Western European countries.

> Concerning the first, a certain literature—which is in contradiction with standard economic theory—suggests that more concentrated industries are responsible for what has been called "administrative inflation": during the upswings of the business cycle, more concentrated industries would use their monopolistic power to increase selling prices more than would be justified by increases in demand or costs. Our analysis of Belgian, Dutch and French wholesale prices over the period 1958–65 shows very clearly that this was not the case. Over a period, more concentrated industries behaved exactly like less concentrated industries. [pp. 172–73]

[2] See also Means (1939), p. 143. John Blair emphasized the same point in defending Means' position: "Stigler implies that the basis of Means' classification was frequency of change—a basis which he then goes on to characterize, in the first of a series of colorful adjectives, as 'whimsical'. Actually, the primary basis was concentration . . ." (p. 70).

[3] The full version of the administered-price thesis deals with the movement of administered prices relative to competitive prices. The tests in Table 1 are performed for a more restricted thesis, labeled by Means as the "truncated" version, which considers the cyclical direction of absolute changes in prices.

TABLE 1—TESTS OF THE ADMINISTERED-PRICE HYPOTHESIS FOR 63 *NB* PRICE INDEXES DURING JULY 1957 TO MARCH 1962, AND 90 *WPI* INDUSTRY SECTOR PRICE INDEXES FOR NOVEMBER 1969 TO NOVEMBER 1970

	Concentration level[a]				
	Less than 40 Percent	40 to 60 Percent	Over 60 Percent	Chi-Square Value[b]	Degrees of Freedom
	$n=16$	$n=29$	$n=18$		
1. Contraction 7/57–4/58				3.12	2
Increase or No Change	37.5%	58.6%	66.7%		
Decrease	62.5	41.4	33.3		
2. Expansion 4/58–6/59				2.47	2
Decrease or No Change	56.2	62.1	38.9		
Increase	43.8	37.9	61.1		
3. Contraction 1/60–1/61				0.42	2
Increase or No Change	50.0	44.8	38.9		
Decrease	50.0	55.2	61.1		
4. Expansion 1/61–3/62				1.15	2
Decrease or No Change	75.0	79.3	88.9		
Increase	25.0	20.7	11.1		
5. Contraction 11/69–11/70	$n=38$	$n=27$	$n=25$	0.16	2
Increase or No Change	81.6	85.2	84.0		
Decrease	18.4	14.8	16.0		
Combined 5 periods				7.32	10

[a] Four-firm concentration for 5-digit SIC product classes; 4-digit SIC industries in panel 5 only.

[b] For 2 degrees of freedom, the chi-square value at the 5 percent level is 5.99; at the 1 percent level 9.21. For 10 degrees of freedom, the respective values are 18.31 and 23.91.

Sources: Stigler and Kindahl; U.S. Bureau of the Census (1966, 1970), and U.S. Bureau of Labor Statistics.

square value well below the level required for significance.

The most recent contraction, November 1969–November 1970, provides an additional opportunity to test the administered-price thesis, using the Wholesale Price Index (*WPI*) industry sector price indexes now available for 4-digit SIC industries. The data in panel 5 of Table 1 show no significant differences between the groups with respect to the percentages of industries in which prices increased during the downturn. Shifting from analysis of directions to magnitudes of change, Means states that during the 1970 recession "... the *BLS* index of industrial prices has risen substantially and administration-dominated prices have risen even more" (1972, p. 292), but he presents no supporting data. The average price increases of 2.7, 5.1, and 3.6 percent, for low, medium, and high concentration industry groups, respectively, do not support Means' assertion.[4]

TABLE 2—FREQUENCY OF PRICE CHANGE FOR 50 ADMINISTERED PRICES, RELATIVE TO COMPETITIVE PRICES, JULY 1957 TO MARCH 1962

	7/57–4/58 Contraction	4/58–6/59 Expansion	1/60–1/61 Contraction	1/61–3/62 Expansion
Decrease	45	13	45	43
No Change	1	8	3	6
Increase	4	29	2	1
	50	50	50	50

Source: See Table 1.

We developed several variations of the analysis to determine whether the tests might be sensitive to the classification systems utilized. We also performed the analysis combining the two contractions, the two expansions, and the two cycles. None of the chi-square results were close to being significant.[5]

---

[4] The difference between the lowest and the midrange of concentration is statistically significant, but the difference between the lowest and the highest concentration group is not. Further, the difference between a less than 50 percent concentration group and an over 50 percent concentration group is not significant.

Means argues that Stigler and Kindahl inappropriately tested only a truncated version of his administered-price thesis. He emphasizes that the administered-price thesis focuses on *relative* price behavior. We therefore tested the full version in Table 2. Additionally, since the results in Table 1 may be influenced by our use of Means' concentration criterion for identifying administered-price industries, we adopted his present classification of 50 administered-price industries in the Table 2 tests.[6] For the competitive price deflator required, we used the *WPI* for commodity groups which Means had earlier identified as competitively priced during his testimony before the Kefauver Committee in 1958 (1959, p. 4752). Those identified as competitive were processed food, hides and leather, lumber and wood, textiles and apparel, and farm products. The average price changes in these five categories were used as the deflator to the National Bureau (*NB*) (Stigler-Kindahl) price series.

The relative price movements of the 50 administered prices designated by Means are shown in Table 2. During both contractions, 90 percent of the relative price changes were decreases, overwhelmingly in disagreement with the Means' thesis. During expansion one, increases took place in 58 percent of the commodities, which also contradicts the thesis. During expansion two, the relative increases were only 2 percent. Thus, Table 2 demonstrates that administered prices were generally more cyclically flexible than the average of competitive prices.

Summarizing, we test the truncated version of the administered-price thesis in Table 1, using Means' earlier criterion of concentration to group industries. In Table 2, we adopt his current designation of administered-price industries to test the full version. Our findings are consistent with neoclassical theory (Stigler, p. 8) in denying the administered-price thesis.

REFERENCES

J. M. Blair, "Administered Prices and Oligopolistic Inflation: A Reply," *J. Bus., Univ. Chicago,* Jan. 1964, *37,* 68–81.

G. C. Means, *Hearings on Administered Prices, Part 9,* Senate Subcommittee on Antitrust and Monopoly, Washington 1959, 4745–60.

———, *Hearings on Economic Concentration, Part 1,* Senate Subcommittee on Antitrust and Monopoly, Washington 1964, 22.

———, "The Administered-Price Thesis Reconfirmed," *Amer. Econ. Rev.,* June 1972, *62,* 292–306.

L. Phlips, *Effects of Industrial Concentration: A Cross-Section Analysis for the Common Market,* Amsterdam 1971.

G. J. Stigler, "Administered Prices and Oligopolistic Inflation," *J. Bus., Univ. Chicago,* Jan. 1962, *35,* 1–13.

——— and J. K. Kindahl, *The Behavior of Industrial Prices,* New York 1970.

National Resources Committee, *The Structure of the American Economy,* Washington 1939.

U.S. Bureau of the Census, Subcommittee on Antitrust and Monopoly, Committee on the Judiciary, U.S. Senate, *Concentration Ratios in Manufacturing Industry, 1963,* Part I, Table 4, Washington 1966.

———, Census of Manufactures, 1967, *Concentration Ratios in Manufacturing, MC67 (s)-2.1,* Washington 1970.

U.S. Bureau of Labor Statistics, *Wholesale Prices and Price Indexes,* Washington.

U.S. Department of Justice, "Merger Guidelines," May 30, 1968, 1–27.

---

[5] Appendix tables containing these additional results are available upon request.

[6] Because the 1970 contraction data in Table 1 were taken from the *BLS* Industry Sector Prices comprising a different set of industries than the Stigler-Kindahl series, in Table 2 we were limited to the time periods covered in Means' article.

Job Quitting and Frictional Unemployment

By J. PETER MATTILA*

The purpose of this note is to point out that much *voluntary* job changing never contributes to measured unemployment or to what has commonly been termed frictional unemployment. Data, which I will present, indicate that at least 50 to 60 percent of all workers line up their new jobs before quitting and leaving their old jobs. This has important implications for the microeconomic work of Armen Alchian, Charles Holt, Dale Mortensen, Edmund Phelps and others on employment, unemployment, and inflation in labor markets. Much of their work has been motivated by the Phillips curve phenomenon. This paper also has important implications for a general understanding of job search behavior and mobility in labor markets.

James Tobin, in his 1971 presidential address, has pointed out that:

> A crucial assumption of the [new microeconomic] theory is that search is significantly more efficient when the searcher is unemployed, *but almost no evidence has been advanced on this point.* Members of our own profession are adept at seeking and finding new jobs without first leaving their old ones or abandoning not-in-labor-force status. [p. 6, italics added]

Tobin goes on to mention the studies of blue-collar worker mobility conducted by Lloyd Reynolds, Herbert Parnes, and others in the late 1940's and early 1950's. They concluded that relatively few workers had a job lined up before quitting and that most job search was conducted while unemployed.

More recent evidence, however, suggests that these conclusions must be modified. Tobin was correct in questioning the conclusions of Reynolds and others and the assumptions of some of the new microeconomic theory.

I

I have presented a model of quit behavior in my 1969 article in which I argue that a worker maximizes utility by lining up a job before quitting whenever the expected costs of unemployed search exceed the additional payoff of a more intensive search while unemployed.

In a recession, it is clear that the costs of unemployed search rise as job vacancies disappear and as the duration of unemployment rises. These rising costs induce potential quitters to be more cautious and reduce the frequency of quits. It is worth noting that since 1967, when the Census Bureau began collecting data on the reason for unemployment, the ratio to the civilian labor force of persons unemployed because they quit their last job has remained steady at 0.4 percent until 1971 when it rose to 0.5 percent for males over 20 years of age.[1] This stability occurred in the face of the 1970–71 recession which almost doubled the unemployment rate for males (over 20 years of age) who were laid off or discharged. The small increase that did occur for quitters is due in part to the increase in the duration of unemployment.[2]

My model also helps us understand why, as Tobin noted, economists typically line up a new job before quitting. The high cost of foregone income while unemployed undoubtedly is a deterrent. An economist typically has greater bargaining power if currently employed than he does without any options to fall back upon. In addition, the national scope of the market makes necessary the relatively cheap methods of preliminary screening (telephone and written

---

* Assistant professor of economics, Iowa State University. I wish to thank Charles Holt, Martin David, and Herbert Parnes for stimulating my interest in this topic and for comments on various versions of this paper.

[1] The unemployed quit rate for females (20+) rose from 0.6 percent in 1968–69 to 0.8 percent in 1970–71 while the teenager rate rose from 1.5 to 1.7 percent in 1970 dropping to 1.6 percent in 1971. (U.S. Dept. Labor, pp. 182–84).

[2] In 1963 only 43 percent of unemployed male quitters experienced 5 weeks or more unemployment as compared to 58 percent in 1971 (U.S. Dept. Labor, pp. 182–84).

communication, the convention "slave market") which are almost as easily executed while employed as unemployed. Of course, other factors which are peculiar to our profession, such as the "April 15th rule" and advance contracting for the academic year in education, undoubtedly play a role as well.

Many of these factors are not unique to the economics profession but are also important in many of the professional, technical, and managerial occupations which now employ over 25 percent of the labor force. Other white-collar and skilled blue-collar workers probably also find that bargaining and income advantages occur when they line a job up in advance. They may not be aware of such advantages when they are young and make their first job change, but they learn by experience.

Albert Rees and George Shultz in studying the search behavior of workers in Chicago concluded that white-collar workers such as typists, keypunch operators, tabulating machine operators, and accountants do in fact rely heavily upon "formal" methods of search (private employment agencies, training school agencies, newspaper ads) whereas blue-collar workers typically found jobs by "informal" methods (gate applications, referral by a friend). They go on to report that:

> For many of the voluntary movers, a new job has been secured before the worker has stopped work on the old one, and the number of months out of work is zero. The occupational group that is the exception to this pattern is typists and keypunch operators, where the probability is highest that the voluntary moves include some temporary withdrawals from the labor force.
> [p. 215]

Unfortunately, they do not publish the detailed data upon which these statements are based.

John Lansing and Eva Mueller (pp. 211–15) confirm that workers who relocate geographically usually line a job up before moving. In a national sample of geographic movers who went to work for a new employer (excluding job transfers), 64 percent reported that they had prearranged their job before moving. This probability increased with educational level and occupational status. A full 85 percent of professional and technical workers versus 61 percent of other white-collar and 49 percent of blue-collar workers secured a job in advance.

By focusing upon relatively low wage blue-collar workers in local labor markets, the early studies of Reynolds, Parnes, and others missed the more systematic and cautious job search patterns which are characteristic of a growing portion of the labor force.[3]

II

In this section, I have collected estimates of the proportion of quitters who line up a job in advance. Much of the data have been published elsewhere but are summarized and evaluated for reliability here. In addition, new estimates based upon previously unpublished data from two national surveys are presented for the first time.

Table 1 summarizes these studies. The estimates of the proportion of quitters who line up jobs in advance varies widely from 31 to 85 percent. But as may be expected, the quality and interpretation of the estimates also varies.

The four national surveys listed at the top of Table 1 are most general in that all occupations and regions were surveyed. The other five studies interviewed factory workers in a particular small or medium size city or region. Unfortunately each of the national surveys also has one or more biases which in part offset their advantage.

The major problem with three of the national samples (the 1955 and 1961 U.S. Bureau of Labor Statistics (*BLS*) surveys (1959, 1963) of job mobility and the Longitudinal Labor Market Survey (see Parnes et al.)) is that estimates have generally been available only for the subset of quitters who changed jobs *once and only once* during the year preceding the interview. For instance, in 1961 the *BLS* (1963, p. 4) reports that 80 percent of those who quit once to improve job status experienced no unemployment be-

---

[3] For a related analysis, in a broader context, see Parnes (1970, pp. 54–56).

TABLE 1—PROPORTION LINING UP A JOB BEFORE QUITTING

Survey	Scope	Period Covered	Number of Quitters in Sample	Percent Lining Up Job
BLS (1963, p. 4)	National	1961	500+[g]	53–65[a] (80[b])
BLS (1959, p. 39)	National	1955	500+[g]	71[b]
Longitudinal	National			
	Men 45–59	1966–67	179	66[b,h]
	White-Collar			77[b,h]
	Blue-Collar			60[b,h]
Lansing-Mueller (pp. 211, 215)	National	1958–63	400	64[c]
	Prof.-Tech.			85[c]
	Other White-Collar			61[c]
	Blue-Collar			49[c]
	Factory Workers Only			
Adams-Aronson (p. 35)	Auburn, N.Y.	1949–50	53	81
Myers-Schultz (pp. 28–30)	Nashua, N.Y.	1948	48	57[d,f]
Reynolds-Gregory (1965, p. 232)	Puerto Rico	1945–54	273[g]	
	Men			50[e]
	Women			38[e]
Reynolds (1951, p. 104)	New Haven, Conn.	1945–47	112[g]	38[e]
Wilcock-Sobel (pp. 63–64)	Kankakee, Ill.	1951–52	108	31[f]

[a] Estimated by author. See text and fn. 5.
[b] Overestimate due to restriction to those changing jobs only once.
[c] Not strictly comparable to other estimates. Refers to geographic movers.
[d] Adjusted by author for 12 percent of quits who left labor force.
[e] Underestimate due to inclusion of quits to leave labor force.
[f] Underestimate due to survey techniques as discussed in text.
[g] Estimated by author.
[h] Estimated by author. See text and fn. 6 and 7.

tween jobs.[4] Persons changing jobs two or more times have been excluded. To the extent that multiple job changers are more likely to be impetuous and to be "risk takers" than those changing jobs only once, the published data overstates the true propensity to line up a job.

Fortunately, additional unpublished data on these multiple job quitters is available for 1961.[5] Using these figures, I estimate that even if *none* of the multiple quitters ever lined up a job, approximately 41 percent of *all* quits would have avoided unemployment. More realistically, if between 25 and 50 percent of all multiple changers obtained a job in advance, the total propensity in 1961 would lie between 53 and 65 percent. These are the figures reported in Table 1 for the 1961 *BLS* survey of job mobility. Unfortunately, such adjustments cannot be made for the 1955 *BLS* survey nor for the Longitudinal Survey. Hence, their results must be interpreted as overestimates. In addition, other problems plague these two surveys which make inference difficult.[6]

---

[4] Each quitter was asked, "Did (he/she) move directly from his first job into this (next) job without losing any time?" (U.S. Census Bureau). Most of the other surveys also inquired about unemployment between jobs. Only two (Reynolds-Gregory, p. 327, and Lansing-Mueller, p. 373) asked explicitly whether the job was lined up in advance. Reynolds-Gregory, for instance, inquired: "Had you already obtained employment at (refer to new job) when you left this one?"

[5] Of the 2,992,000 people making one or more quits to improve job status in 1961 2,225,000 made only one such change. The other 767,000 workers made 2,052,000 such quits, an average of 2.67 quits per person. I am grateful to Harold Goldstein and the *BLS* for supplying this data.

[6] Workers were asked if they experienced any unem-

Lansing and Mueller's estimates are not strictly comparable to the others in Table 1. They asked if a job was lined up prior to a geographic move. The worker could have been laid off, fired, or quit his previous job. Yet because of the unique information which they provide about the geographically mobile, their estimates are included.

Some other problems of interpretation plague the surveys of factory workers. Reynolds, pp. 101–05, and Reynolds and Peter Gregory, p. 232, were not able to exclude workers who quit to leave the labor force. This is a serious problem with respect to women who often quit to return to household work. But in addition some men may quit because of school or illness. It is clear that these two studies underestimate the statistic which we are trying to measure. Two other studies (see Richard Wilcock and Irvin Sobel, pp. 63–64, and Charles Myers and Shultz, pp. 28–30, 205) did not inquire specifically whether the worker had lined up a job, but apparently, if the information was volunteered, it was recorded. Hence, they also yield underestimates.

Weighing the quality of the various surveys, I conclude that the adjusted 1961 *BLS* estimates are probably the most reliable measures available today.[7] At least one-half and perhaps two-thirds of all quitters secure a job prior to quitting. Even among the six estimates available from the factory worker surveys, three report probabilities of 50 percent or higher. Furthermore, there is substantial evidence that white-collar workers, particularly professionals and managers, are far more likely to line up a job than are blue-collar workers.

III

In this section I briefly summarize the available data on the wage rate changes reported by quitters. Reynolds, p. 215, reports that 60 percent of those who lined up jobs improved their wage versus 25 percent of those who didn't secure a job in advance. Parnes (1954, p. 176–77), in reviewing the early studies, concluded that 50–60 percent of all quitters achieved an immediate financial gain.

More recently, Lansing and Mueller, p. 249 have estimated that 72 percent of those prearranging a job before a geographic move secured higher pay compared to 55 percent of those who didn't prearrange a job in advance. Rees and Schultz, pp. 212–14, report that 57 percent of all quitters in their sample obtained higher wages.

Turning to the Longitudinal Labor Market data, approximately 69 percent of those who reported 52 weeks of employment achieved higher hourly wages on the second job in 1966–67 and had a median wage increase of 10.7 percent.[8] On the other hand, only 51 percent of those reporting some unemployment between jobs (excluding those who quit due to bad health) received higher wages so that their median wage change was barely above zero.

It is clear that a small but positive proportion of quitters who line up a job in advance take a wage cut on the new job. Some, without doubt, may quit and move only to find that the job wasn't available or wasn't as attractive as first believed. Most, however, probably give up some pay in order to obtain better working conditions, fringe benefits, chances for advancement, and other factors not reflected in hourly pay data.

IV

In conclusion, the job search behavior of economists does not appear to be unique. At least 50 to 60 percent of all quits move from job to job without ever experiencing unem-

---

ployment during the year and this was attributed to the job change. For instance, men in the Longitudinal Survey reported that only 16 percent had experienced one or more weeks of unemployment, another 18 percent reported spending some time out of the labor force, while 66 percent reported working 52 weeks during the year. The questionnaire was too imprecise to determine whether time was lost between jobs. This was not a problem in the 1961 survey. I am grateful to Herbert Parnes for making the Longitudinal data available. For further information see Parnes et al.

[7] Within a few years even better data will become available from the Longitudinal Labor Market Survey. In the 1971 reinterview, workers in all cohorts (men 45–59; women 30–44; boys and girls 14–24) have been asked whether they had obtained a job in advance of quitting another job.

[8] Hourly wage data for one or both jobs was not available for 29 percent of the observations. For further information see Parnes et al.

ployment. It is rational search strategy to line up a job in advance of quitting in order to avoid the costs of foregone income and to maximize bargaining power. This is a very efficient search strategy for high income people in national or regional labor markets. And as incomes rise and labor markets broaden, we can probably expect increasing use of such search strategy.

I recommend that in future model building, quit rates be viewed as having two components: a fairly small, constant exogenous flow of quits into unemployment (those quitting impetuously after a fight with the boss) and a larger, endogenous flow of utility maximizing workers who move directly from job to job.

This observation gives us many insights. For instance, it explains the paradox that in the 1970–71 recession quit rates fell rapidly while quit-caused unemployment was stable or rose slightly (U.S. Dept. Labor, pp. 182–84, 225). As job opportunities grow in a period of economic expansion, quit rates increase rapidly and quit-caused unemployment should be steady or drop slightly.

It emphasizes that those who argue (rightly) that some unemployment is necessary in an economy if we are to have an efficient allocation of resources, are referring primarily to unemployment resulting from employer initiated lay offs and not to voluntary quits.[9] The strong political distaste for unemployment of any type is not surprising in view of workers' "revealed preference" for employed search before quitting.

Finally, it suggests that the new microeconomic theories (see Phelps et al.) which assume that quitters must pass through a period of unemployment, should be reexamined. As Tobin has emphasized, their models and inferences " . . . would be more convincing if they were corroborated by direct evidence" (p. 8).

REFERENCES

L. P. Adams and R. L. Aronson, *Workers and Industrial Change,* Ithaca 1957.

C. C. Holt and M. H. David, "The Concept of Job Vacancies in a Dynamic Theory of the Labor Market," in *The Measurement and Interpretation of Job Vacancies,* Nat. Bur. Econ. Res. conference report, New York 1966.

J. B. Lansing and E. Mueller, *The Geographic Mobility of Labor,* Ann Arbor 1967.

J. P. Mattila, (1969a) "Quit Behavior in the Labor Market," *Amer. Statist. Assn. Proc.,* Aug. 1969, Econ. Statist. Section, 697–701.

———, (1969b) "Theory and Estimation of Quit Functions," doctoral dissertation, Univ. Wisconsin 1969.

C. A. Myers and G. P. Shultz, *Dynamics of a Labor Market,* New York 1951.

H. S. Parnes, *Research on Labor Mobility,* New York 1954.

———, "Labor Force Participation and Labor Mobility," in *A Review of Industrial Relations Research Vol. I,* Ind. Rel. Res. Assn. report, Madison 1970.

——— et al., *The Pre-Retirement Years Vol. 2,* Center for Human Resource Research report, Columbus 1970.

E. S. Phelps et al., *Microeconomic Foundations of Employment and Inflation Theory,* New York 1970.

A. Rees and G. P. Shultz, *Workers and Wages in an Urban Labor Market,* Chicago 1970.

L. G. Reynolds, *The Structure of Labor Markets,* New York 1951.

——— and P. Gregory, *Wages, Productivity, and Industrialization in Puerto Rico,* Homewood 1965.

J. Tobin, "Inflation and Unemployment," *Amer. Econ. Rev.,* Mar. 1972, *62,* 1–18.

R. C. Wilcock and I. S. Sobel, *Small City Job Markets,* Urbana 1958.

U.S. Bureau of the Census, "Survey of Job Experience in 1961," Interview Form CPS 514, Washington 1961.

U.S. Bureau of Labor Statistics, *The Extent and Nature of Frictional Unemployment,* Study Paper No. 6 for Joint Economic Committee of U.S. Congress, Washington 1959.

———, "Job Mobility in 1961," Special Labor Force Report No. 35, Washington 1963.

U.S. Department of Labor, *Manpower Report of the President,* Washington 1972.

---

[9] They would also be correct in referring to the need for some unemployment of inflowing housewives and students into the labor force. New entrants need time to gather information about available jobs. On the other hand, many college students appear to line up jobs before finishing school. In 1971, 46 percent of the unemployed had involuntarily lost their last job, 42 percent had entered or reentered the labor force, and only 12 percent had quit a job (U.S. Dept. Labor, p. 184).

Soviet Postwar Economic Growth and Capital-Labor Substitution: Comment

By T. KRISHNA KUMAR AND EPHRAIM ASHER*

In a recent article in this *Review*, Martin Weitzman attempted to estimate an aggregate production function for Soviet industry and subsequently explain its postwar pattern of growth. In doing so he specified the following Constant Elasticity of Substitution (*CES*) production function with disembodied Hicks-neutral technical progress

$$(1)\quad Y(t) = \gamma e^{\lambda t}[\delta K(t)^{-\rho} + (1-\delta)L(t)^{-\rho}]^{-1/\rho}u(t)$$

where $u(t)$ is a multiplicative disturbance term having a lognormal distribution with mean unity. Weitzman estimated its parameters using a non-linear regression program after taking logarithmic transform. While the use of this technique is indeed preferable to the linear regression technique (based on profit maximizing marginal conditions) in its applicability to a nonmarket oriented economy, such as that of Russia, it may also lead to erroneous conclusions. In its blessings may lie a dangerous booby trap awaiting the unaware empiricist. It is the purpose of this comment to show a main shortcoming associated with this approach and its impact on empirical research on production.

I

Economists using linear regression models and the usual least squares estimation technique end up with a single set of parameter estimates. They base their empirical conclusions on such a set subject, of course, to the usual set of statistical and theoretical constraints. The use of non-linear least squares programs may result, on the other hand, in *multiple* sets of parameter estimates.

Weitzman asserted in a footnote that "... While this algorithm can at best obtain only a *local* minimum, converging to the same parameter values with different initial estimates makes me confident that the estimates given here in fact globally minimize the error sum of squared residuals" (p. 680). This statement, while hiding more than it reveals, could also mislead some to believe that in similar situations we ought to accept those estimates that give the least possible value to the error sum of squares (*ESS*). Weitzman either experimented within a narrow neighborhood of his optimal parameter estimates always ending up at the same point which could very well be the one corresponding to the global minimum, or he experimented with a variety of initial parameter values that could have generated different local minimum values of the *ESS*. In the latter case, he did not provide the reader with a chance to look at the other minima probably on the ground that the *ESS* at the other local minima were larger than the *ESS* associated with his parameter estimates. Table 1 shows some other local minima of the *ESS* in addition to the one obtained by Weitzman (Case 7).[1] We reached these using the same set of data, and the same method of estimation as the ones used by Weitzman but with different sets of *initial* parameter values. Using a tolerance of 0.00001 we obtained alternative parameter estimates corresponding to different local minima of the *ESS*.

An important observation can be made by looking at Table 1. There exist significantly

---

* Associate and assistant professors, Florida State University, respectively. We wish to thank our colleagues John Chang and James Gapinski for their valuable comments. We are especially grateful to the anonymous referee who made a very perceptive comment on the original results.

[1] In estimating the parameters in Table 1 we have assigned t=1 for the year 1950, while Weitzman assigned t=0 for the same year. This is the reason for the slight difference between our estimate of γ in Case 7 and that of Weitzman.

TABLE 1—ALTERNATIVE FINAL PARAMETER ESTIMATES ASSOCIATED WITH ALTERNATIVE INITIAL PARAMETER VALUES[a,b]

Cases	γ	λ	δ	σ	R^2
1					
Initial Values	1.2500	0.0500	1.0000	1.2500	0.9871
Final Estimates	0.8759	0.0112	0.8151	1.0257×10^{11}	0.9949
	(0.2645)	(0.0265)	(0.3438)	(2.0846×10^{22})	
2					
Initial Values	1.2500	0.0500	1.0000	1.10000	0.9881
Final Estimates	0.8504	0.0133	0.7966	2.6470×10^{10}	0.9945
	(0.2481)	(0.0256)	(0.3324)	(1.2216×10^{21})	
3					
Initial Values	1.0000	0.3000×10^{-5}	0.9999	0.9999	0.9906
Final Estimates	1.0000	0.1676×10^{-8}	0.9998	0.1993	0.9980
	(0.0899)	(0.0094)	(0.1504)	(68.6390)	
4					
Initial Values	0.2500	0.0300	0.2500	1.2500	0.999512
Final Estimates	0.7858	0.0207	0.6368	0.4024	0.999513
	(0.0468)	(0.0053)	(0.0701)	(0.0300)	
5					
Initial Values	0.9990	0.2500×10^{-4}	0.9900	0.9990	0.9918
Final Estimates	0.9989	0.2144×10^{-8}	0.9893	0.9014	0.9980
	(0.3016)	(0.0270)	(0.3371)	(12.6430)	
6					
Initial Values	0.2500	0.0050	0.7500	5.0000	0.999509
Final Estimates	0.7371	0.0205	0.6409	0.4042	0.999514
	(0.0472)	(0.0054)	(0.0706)	(0.0306)	
7					
Initial Values	0.5000	0.1000×10^{-5}	0.6500	5.0000	0.999513
Final Estimates	0.7876	0.0205	0.6390	0.4030	0.999513
	(0.0469)	(0.0053)	(0.0702)	(0.0302)	

[a] The numbers in parentheses are the standard errors of the estimated coefficients based on a linearized regression about the final parameter values.

[b] The table gives two alternative R^2s for each case. The top R^2 is the commonly used proportion of sum of squares explained by the regression. The second R^2 is the square of the correlation between the actual and predicted values of the dependent variable.

different sets of parameter estimates obtained from the same sample data (the data used by Weitzman) that give rise to somewhat insignificant differences in the measure of goodness of fit, R^2. Furthermore, as Cases 4, 6, and 7 show, the sample *ESS* has a somewhat flat bottom at one of its minimum values which *could* very well be the global minimum. In non-linear regression there is no requirement that the mean prediction is equal to the prediction with the sample means. As such the commonly used measure of R^2 (our top figure), proportion of sum of squares explained by regression may not lie in the range between 0 and 1. For this reason we used a second measure R^2 (our bottom figure) whose definition is given in the footnote to Table 1. The differences in R^2 are even smaller with the second measure than with the former.

In situations like this the investigator is faced with a problem of choosing one set of estimates over the others. The footnote of Weitzman referred to earlier seems to suggest that one ought to choose that set of estimates which corresponds to the global minimum of the *ESS* or the global maximum of the top R^2. This could be misleading as the *ESS* could be quite sensitive to sampling fluctuations, errors of measurement, rounding off errors, etc. This question of sensitivity of *ESS* is very crucial when the differences in *ESS* corresponding to different local minima are slight as is suggested by our results. The parameter estimates that correspond to global minimum of *ESS* with one sample may bear no resemblance to the parameter estimates obtained the same way from another sample.

II

To see the implications of the above observations on empirical studies using the direct estimation of the *CES* production, let us examine the results further. The results suggest that the Soviet data could be quite consistent with four entirely different and, indeed, opposing hypotheses. Cases 1 and 2 may be consistent with a *CES* production structure with perfect capital-labor substitution.[2] Case 3, on the other hand, may be consistent with the Leontief structure of production with capital as the limiting factor with virtually no technical progress. The third possibility, given by Case 5, suggests that the data could be emerging from a Cobb-Douglas production structure with virtually no technical progress. Finally, all of these results are different still from the Weitzman hypothesis supported by Cases 4, 6, and 7. Since the differences in R^2s are small between these alternatives the choice between them based only on maximum R^2 could lead the researcher to reject hastily several plausible production structures.

An alternative interpretation of the results in Table 1 is that there exist partial tradeoffs between various parameter estimates as we go from one set of parameter estimates to another with insignificant variations in R^2. This interpretation leads one to suspect that even with a large sample or complete information one may get an *ESS* function which either has a flat bottom (local nonidentification) or multiple global minima (global nonidentification).[3]

What is needed in situations like this is some prior or additional information. The need for such information is now clearly apparent on two grounds: First, in model building that usually precedes estimation, identification is brought about by imposing certain prior restrictions. Second, such information is needed in order to select one set of parameter estimates from various alternative sets associated with multiple minima of *ESS*.[4]

III

Our numerical experiments thus lead us to conclude that in models such as the one considered by Weitzman, the Soviet data may be compatible with a variety of production structures including Leontief and Cobb-Douglas structures. If this in fact is the case, Weitzman's explanations of the postwar pattern of Soviet economic growth given in page 682 (inclusive of Table 3) will no longer be true. This possibility clearly calls for further research on the subject of Soviet economic growth. It also calls for further investigations in the methodology of non-linear model building. Finally we would like to stress that our attempt should serve not just as a caution in accepting Weitzman's results, but also as a caution for those who plan to use similar approaches in empirical economic research.

REFERENCES

F. M. Fisher, *The Identification Problem in Econometrics*, New York 1966.

E. Malinvaud, *Statistical Methods in Econometrics*, 2d ed., Chicago 1970.

T. J. Rothenberg, "Identification in Parametric Models," *Econometrica*, May 1971, *39*, 577–92.

M. L. Weitzman, "Soviet Postwar Economic Growth and Capital Labor Substitution," *Amer. Econ. Rev.*, Sept. 1970, *60*, 676–92.

---

[2] The customary procedure of applying the *t*-ratio for testing the significance of the parameters is, in general, inappropriate in non-linear regression with small samples. Thus, the fact that some of the estimates in Table 1 are associated with relatively large standard errors does not necessarily imply that the parameters are statistically insignificant in any clear way. For further discussion on inference in non-linear regression theory see Edmond Malinvaud.

[3] The question of identification of parameters in stochastic, non-linear models of production has not been investigated to the best of our knowledge. The theoretical treatment of identification in stochastic, non-linear models by Franklin Fisher and its recent extension by Thomas Rothenberg seem to be appropriate for examining this question.

[4] For instance, such information may be provided by the actual factor shares in the economy. Thus the closeness between the imputed and actual factor shares could be used as a criterion for selecting the appropriate set of estimates. However, for the Soviet economy, reliable data on actual shares are currently unavailable to us to make any further comment.

Soviet Postwar Economic Growth and Capital-Labor Substitution: Reply

By MARTIN WEITZMAN*

Because Soviet industrial growth in the postwar period has been very regular, it is easy to get a high R^2 by regressing output on almost anything. For example, running output as a function of capital and exponential technical change, *without* including labor gives a respectible fit. So, for that matter, does running output on exponential technical change alone, without *either* capital *or* labor. T. Krishna Kumar and Ephriam Asher seem to think that the difference between the high R^2 obtained with such forms and the .9997 R^2 obtained under my form of the *CES* production function is insignificant. They offer no statistical theory to buttress their claim. In fact, the appropriate statistic for measuring the significance of such differences is the percentage change in the error sum of squares appropriately corrected for degrees of freedom. By this criterion, my form turns out to be better than any of those proposed by Kumar and Asher at the 95 percent confidence level.

The fact that some particular alternative hypothesis gave a "good fit" because it registered a high R^2 or because it was a local optimum in the error sum of squares minimizing routine is a nonstatistical irrelevancy.

---

* Massachusetts Institute of Technology.

NOTES

FOURTH CONGRESS OF THE INTERNATIONAL ECONOMIC ASSOCIATION

Budapest, Hungary, August 19–24, 1974

ECONOMIC INTEGRATION: WORLDWIDE, REGIONAL, SECTORAL

Monday, August 19, 1974

9:30 A.M. OPENING CEREMONIES
Chairman: Fritz Machlup
Address of Welcome: Jenö Fock, Prime Minister of Hungary

11:00 A.M. *First Plenary Session*
TYPES OF INTEGRATION
Chairman: Raul Prebisch (Argentina)
Main Paper: Bela Balassa (U.S.)
Discussion Papers: Margarita Maksimova (USSR); Richard G. Lipsey (Canada)

3:00 P.M. *Second Plenary Session, First Half*
WORLDWIDE VERSUS REGIONAL INTEGRATION: IS THERE AN OPTIMUM SIZE OF THE INTEGRATED AREA?
Chairman: Tigran Khachaturov (USSR)
Main Paper: Richard N. Cooper (U.S.)
Discussion Papers: Germánico Salgado (Ecuador); Evgueni Mateëv (Bulgaria)

5:00 P.M. *Second Plenary Session, Second Half*
PRESIDENTIAL ADDRESS
Chairman: G. Ugo Papi (Italy)
The History of Thought on Economic Integration: Fritz Machlup, President

Tuesday, August 20, 1974

9:00 A.M. *Third Plenary Session, First Half*
SOCIO-POLITICAL AND INSTITUTIONAL ASPECTS OF INTEGRATION
Chairman: Robert K. A. Gardiner (Ghana)
Main Paper: Joseph Bognar (Hungary)
Discussion Papers: Robert Marjolin (France); Roman Moldovan (Rumania)
(Discussion to be continued in Working Group G)

11:00 A.M. *Third Plenary Session, Second Half*
MEASURING THE DEGREE OR PROGRESS OF INTEGRATION
Chairman: Nicolas Kaldor (U.K.)
Main Paper: Jean Walbroeck (Belgium)
Discussion Papers: Willi Kunz (German Dem. R.); Ricardo Ffrench-Davis (Chile)
(Discussion to be continued in Working Group A)

3:00 P.M. *First Sessions of Working Groups*

Wednesday, August 21, 1974

9:00 A.M. *Second Sessions of Working Groups*
3:00 P.M. *Third Sessions of Working Groups*

Thursday, August 22, 1974

9:00 A.M. *Fourth Sessions of Working Groups*
3:00 P.M. *Fifth Sessions of Working Groups*

Friday, August 23, 1974

FREE FOR EXCURSIONS, RECREATION, AND PREPARATION OF REPORTS

Saturday, August 24, 1974

9:00 A.M. *Fourth Plenary Session*
Chairman: Nurul Islam (Bangladesh)
Reports by chairmen of Working Groups A, B, C, D, and E

3:00 P.M. *Fifth Plenary Session*
Chairman: Poul Nørregaard Rasmussen (Denmark)
Reports by chairmen of Working Groups F, G, H, I, and J

WORKING GROUPS

A. *Measuring the Degree or Progress of Economic Integration*
Chairman: Erik Lundberg (Sweden)
Main Paper: See Third Plenary Session
Discussion Papers: See Third Plenary Session
Opening Statement: Åke Anderssen (Sweden)

B. *Sectoral Integration: Agriculture, Transport, Energy, Selected Industries*
Chairman: Joszef Pajestka (Poland)
Main Paper: Hans Willgerodt (Fed. R. Germany)
Discussion Papers: Zdenek Orliček (Czechoslovakia); Miguel Wionczek (Mexico)
Opening Statement: Victoria Curzon (Switzerland)

C. *Industrial Policy: Location, Technology, Multinational Firms, Competition, and Integration of Product Markets*
Chairman: Harry G. Johnson (U.K.)
Main Paper: Pierre Uri (France)
Discussion Papers: Ryutaro Komiya (Japan); Nino Fidelino de Figuereido (Brazil)
Opening Statement: Peter Veress (Hungary)

D. *Migration and Integration of Labor Markets*
Chairman: Orlando D'Alauro (Italy)
Main Paper: Kosta Mihailović (Yugoslavia)
Discussion Papers: Hilde Wander (Fed. R. Germany); Giorgio Basevi (Italy)
Opening Statement: Preben Munthe (Norway)

E. *International Capital Movements and Integration of Capital Markets*
Chairman: Herbert Giersch (Fed. R. Germany)
Main Paper: Peter Kenen (U.S.)
Discussion Papers: Robert Mundell (Canada); Alexander Swoboda (Switzerland)
Opening Statement: Raymond Bertrand (France)

F. *Monetary and Fiscal Integration*
Chairman: Raymond Barre (France)
Main Paper: Alexandre Lamfalussy (Belgium)
Discussion Papers: Carlos Massad (Chile); John Williamson (U.K.)
Opening Statement: Marina von Neumann Whitman (U.S.)

G. *Socio-Political and Institutional Aspects of Integration*
Chairman: Shigeto Tsuru (Japan)
Main Paper: See Third Plenary Session
Discussion Papers: See Third Plenary Session
Opening Statement: Emil Salim (Indonesia)

H. *Integration of Less Developed Areas and of Areas on Different Levels of Development*
Chairman: Indraprasad G. Patel (India)
Main Paper: Eduardo Lizano (Costa Rica)
Discussion Papers: Geoffrey Denton (U.K.); Mahbub ul Haq (Pakistan); Samir Amin (Egypt); Gunal Kansu (Turkey); Dharam Ghai (Kenya)
Opening Statement: Helen Hughes (Australia)

I. *Integration by Market Forces and Through Planning*
Chairman: Béla Csikós-Nagy (Hungary)
Main Paper: Oleg T. Bogomolov (USSR)
Discussion Papers: Gottfried Haberler (U.S.); Witold Trzeciakowski and Jerzy Mycielski (Poland)
Opening Statement: Harry Edward English (Canada)

J. *World Trade and Intraregional Trade: Trends and Structural Changes*
Chairman: Petrus J. Verdoorn (Netherlands)
Main Paper: Gunther Kohlmey (German Dem. R.)
Discussion Papers: Nadav Halevi (Israel); Mikhail V. Senin (USSR)
Opening Statement: M. G. Mathur (India)

REGISTRATION: Inquiries and requests for application forms and hotel reservations may be sent to the International Economic Association, 54, Boulevard Raspail, 75270 Paris Cedex 06, France, or to the Bureau of the Fourth World Congress of Economists, Kossuth Lajos ter. 11, 1860 Budapest 55, Hungary. The registration fee is Swiss francs 90 if paid by March 31; S.f. 120 if paid before July; of S.f. 150 if paid later or in Budapest. Payments can be made in all convertible currencies at the current exchange rates.

The following Statement of Ownership, Management, and Circulation is provided in accordance with the Act of October 23, 1962; Section 4369, Title 39, United States Code. The *American Economic Review* is owned and managed by the American Economic Association, a non-profit educational organization, and is printed in Menasha, Wisconsin, by the George Banta Company. During the preceding 12 months the average number of copies printed for each issue was 26,600 the average paid circulation, 22,914; the average free distribution, 282, the average number of copies distributed, 23,196. Corresponding figures for last issue before filing: 26,000 copies printed; 22,994 paid circulation; 280 free copies distributed; 23,276 total number of copies distributed.

ANNOUNCEMENTS

Starting January 1, 1974, all manuscripts submitted to this *Review* have been sent out to referees without the author's name or affiliation. Original submissions should have a separate cover page for the title, author's name, and affiliation. The first page of the text should carry the title of the paper. For purposes of refereeing, authors are requested to avoid self-identification in the main text of their papers.

The P.W.S. Andrews memorial prize is awarded annually for an essay by a young scholar (under the age of 30 or within eight years of taking his first degree) in the general field of Industrial Economics and the Theory of the Firm, broadly interpreted. The prize is £100 (or the equivalent in other currency) and the winning essay will normally be published in *The Journal of Industrial Economics*. An essay submitted should be a work of original research by the candidate only, not previously published, and not previously awarded any other prize. It should be submitted in English and should not exceed 10,000 words in length. The next closing date for entries is December 31, 1974. Candidates should obtain details of the conditions of entry from the Academic Registrar, University of Lancaster, Bailrigg, Lancaster, England.

The Faculty Exchange Center (FEC) aims to make it possible for an interested faculty member to exchange position with a colleague from another institution either on this continent or overseas where instruction is in English. The Center will bring to the attention of its members the names of others interested in exchange through the publication of a catalog containing the names of the instructors and their institutions, their fields of specialization, and the region where they would like to teach. The catalog will be made available to registered members only. The first FEC catalog showed a membership representing over 30 disciplines from over 100 colleges and universities from the United States, Canada, and seven countries overseas. For more information, write for FEC brochure and registration forms: Box 1866, Lancaster, Pennsylvania 17604. Both appear on outer cover of the September A.A.U.P. Bulletin. You may use the registration form appearing there or a copy of it.

New Journal

The department of peace science of the University of Pennsylvania, in collaboration with the Peace Science Society (International), has begun publication of the *Journal of Peace Science*. The *Journal* is a forum for research toward the underlying theory of conflict and conflict management in its economic, political, psychological, and social dimensions. The *Journal* will publish studies of conflict and peace theory and scientific investigations into conflict management and resolution. Manuscripts may be submitted to the Editors, *Journal of Peace Science*, McNeil Building, 3718 Locust Walk, University of Pennsylvania, Philadelphia, Pa. 19174. Prospective contributors are invited to write for a style sheet.

The Czechoslovak Society of Arts and Sciences in America will hold its Seventh Congress, November 15–17, 1974 at New York University. Papers dealing with the arts, the humanities, and the natural as well as social sciences will be considered. For details, contact the Program Chairman, Professor Rudolf Sturm, Skidmore College, Saratoga Springs, New York 12866.

The International Institute of Management (IIM) will sponsor an International Congress on Financial Theory and Decision Models. The conference will be held in West Germany in June 1974. Papers in all areas of finance are being sought. However, special consideration will be given to papers developing theory or models that are especially applicable to the European setting. Anyone wishing to attend or present a paper should contact either of the conference cochairmen, Edwin J. Elton or Martin J. Gruber at IIM, 1 Berlin 33, Griegstrasse 5, West Germany.

A Special Summer Program in Urban Economics will be held at Massachusetts Institute of Technology, August 12–23, 1974. Featured will be classroom work and distinguished lecturers discussing the most recent theoretical and empirical economic findings in the area of urban economics ranging from housing and transportation to the provision, organization, and pricing of government services, etc. For further information, please write to Director of the Summer Session, Room E19-356, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

Anyone wishing to contribute a paper to the Medical Care Section Program of the next annual meeting of the American Public Health Association to be held October 20–24, 1974 in New Orleans, Louisiana may obtain standard abstract forms from Donald K. Freeborn, Associate Director, Health Services Research Center, Kaiser Foundation Hospitals, 4707 S.E. Hawthorne Blvd., Portland, Oregon 97215. The deadline for submitting abstracts is March 15, 1974.

On May 6–7, 1974, the Center for Economic Development, University of Texas, is sponsoring a conference on "Alternative Goals for Metropolitan Growth and Development." Papers will be presented by Gordon Cameron, Michael Conroy, John Cumberland, Earl Finkler, Niles Hansen, William Neenan, Harry Richardson, and Wilbur Thompson. The presentations will be followed by workshop discussions emphasizing practical planning approaches to contemporary problems. There will be a registration fee of $50. Inquiries should be addressed to Ms. Rita Ellison, Center for Economic Development, B.E.B. 413, The University of Texas, Austin, Texas, 78712.

The June 1973 issue of this *Review* incorrectly listed Carl Sumner Shoup's middle name in the Distinguished Fellow citation. We apologize for this error and hope that it caused him no embarrassment or inconvenience.

Deaths

Yochanan Comay, senior lecturer and chairman, department of economics, Faculty of Industrial Engineering and Management, Technion-Israel Institute of Technology, Haifa, Israel, Nov. 24, 1973.

Harold Hotelling, Chapel Hill, North Carolina, Dec. 26, 1973.

Retirements

Frank Hanna, department of economics, Duke University, Dec. 1972.

Gault W. Lynn, director of research, Federal Reserve Bank of San Francisco, Apr. 1, 1973.

Visiting Foreign Scholars

Uri Ben-Zion, Technion, Israel: visiting assistant professor in economics, University of Minnesota, June–Sept. 1, 1974.

Mrinal Datta-Chaudhuri, Delhi School of Economics: visiting professor of economics, University of Minnesota, fall and winter quarters, 1973–74.

Rikard Lang, Zagreb University, Yugoslavia: professor of economics, Colgate University, fall 1973.

Tzong-Biau Lin, New Asia College, The Chinese University of Hong Kong: visiting scholar, department of economics, University of Pittsburgh, 1973–74.

Alasdair MacBean, University of Lancaster: visiting professor of economics, University of Michigan.

Yair E. Orgler, Tel Aviv University: visiting scholar, Division of Research, Federal Deposit Insurance Corporation, 1973–74.

Wilfred Prest, University of Melbourne, Australia: visiting professor of economics, University of Pittsburgh.

Assaf Razin, Tel Aviv University: visiting professor in economics, University of Minnesota, Feb.–Sept. 1974.

Jozef Wilczynski, University of New South Wales: visiting professor of economics, University of Pittsburgh, winter 1974.

Promotions

Raveendra N. Batra: professor, department of economics, Southern Methodist University, May 1973.

Vittorio Corbo: associate professor of economics, Sir George Williams University, June 1, 1973.

John G. Cross: professor of economics, University of Michigan, June 1973.

Herman E. Daly: professor, department of economics, Louisiana State University.

James W. Dean: assistant professor, department of economics and commerce, Simon Fraser University.

Joseph Domitrz: associate professor of economics, Western Illinois University, Sept. 18, 1973.

Stephen C. Farber: assistant professor, department of economics, Louisiana State University.

Paul Feldstein: professor of economics, University of Michigan, June 1973.

Stephen E. Guisinger: associate professor department of economics, Southern Methodist University, Feb. 1973.

Kanji Haitani: professor of economics, State University of New York, College at Fredonia, Sept. 1973.

Richard B. Harshbarger: professor of economics, Manchester College, Sept. 1973.

Lamar B. Jones: professor, department of economics, Louisiana State University.

Richard G. Kreitner: assistant director, financial controls and analysis, Group Insurance Operations, Connecticut General Life Insurance Company, Sept. 21, 1973.

Stephen D. Lewis: associate professor of economics, University of Alberta.

Gail E. Makinen: associate professor, department of economics, Wayne State University, Sept. 10, 1973.

Hameed Naz: assistant chief administrative officer, Planning and Monitoring, Montgomery County, Maryland.

Jack Ochs: associate professor, department of economics, University of Pittsburgh.

Michael Parti: assistant professor, department of economics, Louisiana State University.

G. Randolph Rice: associate professor, department of economics, Louisiana State University.

Gerald W. Scully: professor, department of economics, Southern Methodist University, May 1973.

Daniel Singer: associate professor of economics, Western Illinois University, Sept. 18, 1973.

Harlan Smith: professor of economics, University of Minnesota.

Morton Stelcner: associate professor of economics, Sir George Williams University, June 1, 1973.

Ronald Teigen: professor of economics, University of Michigan, Aug. 1972.

Edward Tower: associate professor, department of economics and commerce, Simon Fraser University.

Balder von Hohenbalken: professor of economics, University of Alberta.

Calman R. Winegarden: professor of economics, University of Toledo, Sept. 1973.

Robert N. Wisner: associate professor of economics, Iowa State University, July 1, 1973.

Administrative Appointments

Muriel Armstrong: acting chairman, department of economics; assistant dean, Faculty of Arts, Sir George Williams University, June 1, 1973.

Wilfred L. David, University of Delaware: chairman and professor, department of economics and business administration, Fisk University, Sept. 1, 1973.

Louis C. Gasper, University of Arizona: president and chief economist, Criterion Studies, Inc., Dallas, Texas, Aug. 1973.

Kanji Haitani: chairman, department of economics and business administration, State University of New York, College at Fredonia, Sept. 1973.

H. Robert Heller: acting chairman, department of economics, University of Hawaii, 1973–74.

Allen Kelley: chairman, department of economics, Duke University, Sept. 1, 1973.

Barry D. Rosenfeld: assistant acting chairman, department of economics, Sir George Williams University, June 1, 1973.

Gerhard Rosegger: chairman, department of economics, Case Western Reserve University, Oct. 1, 1973.

Morton Stelcner: assistant acting chairman, department of economics, Sir George Williams University, June 1, 1973.

Yeong-Her Yeh: acting graduate chairman, department of economics, University of Hawaii, 1973–74.

Appointments

Jean W. Adams: assistant professor of economics, Iowa State University.

Michael J. Applegate: visiting assistant professor of economics, Iowa State University, 1973–74.

Hossein G. Askari, Tufts University: associate professor, department of economics, Wayne State University, Sept. 10, 1973.

Burt Barnow, University of Wisconsin: assistant professor, department of economics, University of Pittsburgh.

James T. Barnowe, University of Michigan: visiting assistant professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

Bruce R. Beattie, University of Kentucky: associate professor of economics, Iowa State University, Aug. 1, 1973.

Frederick W. Bell: professor, department of economics, Florida State University, Sept. 15, 1973.

Gerald E. Bennington, North Carolina State University: visiting associate professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

Richard Blacklock: bureau of economics, Federal Trade Commission.

Michael D. Boehlje, Oklahoma State University: associate professor of economics, Iowa State University, Sept. 1, 1973.

John O. Bornhofen: associate professor of economics and finance, F. E. Seidman Graduate College of Business, Grand Valley State Colleges.

Michael Boskin: research associate, National Bureau of Economic Research, Jan. 1, 1974.

Phelim P. Boyle, Duncan C. Fraser & Co. (Actuaries), Liverpool, England: assistant professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

Joanne M. Brasco: bureau of economics, Federal Trade Commission.

Robert D. Brogen: bureau of economics, Federal Trade Commission.

Michael C. Burstein, Northwestern University: visiting assistant professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

Betty Chu, University of Notre Dame: lecturer, California State University, Long Beach, Sept. 4, 1973.

Peter Clark, University of Minnesota: research fellowship, National Bureau of Economic Research, 1973–74.

John D. Claxton, University of Windsor: assistant professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

Paul N. Courant, Princeton University: lecturer in economics, University of Michigan.

Steven R. Cox: bureau of economics, Federal Trade Commission.

Robert W. Crown: assistant professor of economics, Iowa State University, July 1, 1973.

Robert C. Daffenbach, University of Illinois at Urbana: assistant professor, department of economics, Wayne State University, Sept. 10, 1973.

Carl T. Devine, Florida State University: visiting professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

Ronald A. Dulaney, Columbia University: assistant professor of economics, University of Montana, July 1973.

Randyl D. Elkin: assistant professor of economics, West Virginia University, Sept. 1, 1973.

Ray C. Fair: senior research associate, National Bureau of Economic Research, July 1, 1973.

Dale F. Falcinelli: instructor, department of economics, Lafayette College, Sept. 1, 1973.

John A. Farquhar: instructor, department of economics and commerce, Simon Fraser University, Sept. 1, 1973.

Anthony C. Fisher, Resources for the Future, Inc.: associate professor, Bureau of Business and Economic Research and department of economics, University of Maryland, Aug. 17, 1973.

Robert F. Fix: visiting assistant professor of economics, Iowa State University, 1973–74.

Michael S. Fogarty: assistant professor of economics, West Virginia University, Sept. 1, 1973.

Leonard B. Forman: economist, Business Conditions Division, Federal Reserve Bank of New York.

James L. Freund: research associate, National Bureau of Economic Research, Sept. 1, 1973.

Edward J. Frydl: economist, Foreign Research Division, Federal Reserve Bank of New York.

Edwin Fujii: assistant professor of economics, Tulane University, 1973–74.

Alex Gardezky, Temple University: assistant professor, department of economics, Villanova University, Sept. 1, 1973.

Melvin N. Greenball, Ohio State University: visiting professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

Charles K. Harley, University of British Columbia: visiting assistant professor of economics, University of Michigan, fall 1973.

Robert M. Harmel, Jr.: associate professor, department of economics, Iowa State University.

Maury B. Harris: economist, Business Conditions Division, Federal Reserve Bank of New York.

Jerome F. Heavey: instructor, department of economics, Lafayette College, Sept. 1, 1973.

Helen L. Heise: bureau of economics, Federal Trade Commission.

Gary Hewitt, Yale University: visiting assistant professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

Arlene Hoel: economist, Market Statistics Division, Federal Reserve Bank of New York.

Thomas D. Hopkins, Bowdoin College: Irwin Management Co., Columbus, Indiana, Sept. 1973.

Hajime Hori, Brown University: assistant professor of economics, Tulane University, 1973–74.

E. Philip Howrey, University of Pennsylvania: professor of economics, University of Michigan.

Darrel Hueth, University of California, Berkeley: assistant professor of resource economics, University of Rhode Island, Sept. 1973.

F. Thomas Juster, National Bureau of Economic Research: program director, Survey Research Center, Institute for Social Research, University of Michigan, May 1, 1973.

William J. Kahley: assistant professor, department of economics, Lafayette College, Sept. 1, 1973.

Jorge M. Katz: research associate, National Bureau of Economic Research, Sept. 1, 1973.

Robin J. Kendrick: assistant research scientist, Center for Research on Economic Development, University of Michigan and professeur adjoint, Université du Cameroun, Oct. 1973.

Dong-Kun Kim: assistant professor of economics, School of Business Administration, University of Tennessee at Martin, Sept. 1973.

Philip A. Klein: senior research associate, National Bureau of Economic Research, Sept. 1, 1973.

Jan Kmenta, Michigan State University: professor of economics, University of Michigan.

Roland H. Koller II: bureau of economics, Federal Trade Commission.

Alan Kraus, Stanford University: associate professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

Anne O. Krueger, University of Minnesota: visiting professor, Massachusetts Institute of Technology, 1973–74.

Edward B. Leviton: bureau of economics, Federal Trade Commission.

Wesley H. Long: industry economist, U.S. Department of Commerce, Domestic and International Business Administration, Bureau of Competitive Assessment and Business Policy, Office of Business Research and Analysis, June 25, 1973.

James D. Macbeth, University of Chicago: assistant professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

Charlie A. McCormick: bureau of economics, Federal Trade Commission.

James W. McFarland, University of Southwestern Louisiana: assistant professor of resource economics, University of Rhode Island, June 1973.

Peter F. M. McLoughlin, University of New Brunswick: president, Peter McLoughlin Associates Ltd., Economic Consultants.

Joseph P. Magaddino, Virginia Polytechnic Institute and State University: assistant professor, California State University, Long Beach, Sept. 4, 1973.

Vincent J. Malanga: economist, Business Conditions Division, Federal Reserve Bank of New York.

Lewis Mandell: director, Social Science Research Institute and associate professor of economics, University of Maine, Sept. 1973.

Judith Mann: visiting assistant professor of economics, University of Hawaii, 1973–74.

Thomas G. Marx: bureau of economics, Federal Trade Commission.

Marshall H. Medoff: assistant professor of economics, University of Santa Clara, Sept. 1973.

Richard G. Milk, Northeast Louisiana University: associate professor of economics, School of Business, Virginia State College, Sept. 1, 1973.

John J. Miller: research associate, department of economics, Iowa State University.

Richard J. Morrison, Carnegie-Mellon University: assistant professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

William Myslinski, University of Minnesota: instructor, department of economics, Colgate University, Sept. 1973.

Donald A. Norman: bureau of economics, Federal Trade Commission.

Jonathan D. Ogur: bureau of economics, Federal Trade Commission.

Dale K. Osborne, State University of New York at Albany: associate professor of economics, Southern Methodist University, Aug. 1973.

George A. Pascoe: bureau of economics, Federal Trade Commission.

Steven M. Pease, Yale University: assistant professor, department of economics, Wayne State University, Sept. 10, 1973.

Fred G. Pennance, University of Reading: visiting professor, Faculty of Commerce and Business administration, University of British Columbia, fall 1973.

Larry C. Peppers, Eastern Michigan University: manager of economic forecasts, Union Pacific Railroad, Omaha, Nebraska, July 1973.

David C. Petersen: instructor of economics, University of Santa Clara, Sept. 1973.

Joseph A. Plante, U.S. Government: instructor, Rasmussen School of Business, Dec. 1972.

Robert O. Reck: bureau of economics, Federal Trade Commission.

Harry W. Richardson, University of Kent at Canterbury, England: professor, department of economics, University of Pittsburgh.

Marcel K. Richter, University of Minnesota: visiting professor, department of economics, University of California at Berkeley, 1973–74.

Rita M. Rodriguez: research associate, National Bureau of Economic Research, Sept. 1, 1973.

Joseph B. Rose, University of New Brunswick: visiting assistant professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

Samuel M. Rosenblatt: director, Office of Business Research and Analysis, Bureau of Competitive Assessment and Business Policy, Domestic and International Business Administration, U.S. Department of Commerce, Aug. 20, 1973.

John Rutledge: assistant professor of economics, Tulane University, 1973–74.

James A. Seagraves: visiting professor of economics, Iowa State University.

Tae Kun Seo, Texas A&M University: assistant professor of economics, Southern Methodist University, Aug. 1973.

I. Lee Marie Skov, University of Southern California: assistant professor, California State University, Long Beach, Sept. 4, 1973.

Charles F. Stone, Yale University: assistant professor, department of economics, Wayne State University, Sept. 10, 1973.

Roscoe E. Stout: visiting assistant professor of economics, Iowa State University, 1973–74.

Jon Sutinen, University of Washington: research associate, department of resource economics, University of Rhode Island, Sept. 1973.

Margo M. Thorning: bureau of economics, Federal Trade Commission.

Ralph E. Townsend: bureau of economics, Federal Trade Commission.

David G. Tuerck: research economist, Center for International Business, Los Angeles, Sept. 1, 1973.

Aman Ullah: assistant professor of economics, Southern Methodist University, Aug. 1973.

Ferdinand G. Vaandrager, University of Pennsylvania: visiting assistant professor, Faculty of Commerce and Business Administration, University of British Columbia, fall 1973.

John J. Van Belle: economist, Balance of Payments Division, Federal Reserve Bank of New York.

Bernard L. Weinstein, Columbia University: economist, bureau of economics, Federal Trade Commission, Apr. 1973.

Randall D. Weiss, Harvard University: assistant professor, University of Maryland.

Ira B. Whitten: bureau of economics, Federal Trade Commission.

Gail R. Wilensky, Urban Institute: visiting professor of economics, University of Michigan, fall 1973.

Jerome M. Wolgin, Yale University: assistant professor, department of economics, Wayne State University, Sept. 10, 1973.

David Yeater, Cornell University: instructor of economics, Colgate University, Feb. 1974.

Richard Zecher: associate professor of economics, Tulane University, 1973–74.

Donald Zimmerman: bureau of economics, Federal Trade Commission.

Leaves for Special Appointments

William P. Albrecht, University of Iowa: legislative assistant to Senator Dick Clark, Washington, D.C., Jan. 1, 1974.

Robert L. Avinger, Jr., Davidson College: bureau of economics, Federal Trade Commission, Aug. 1973–Aug. 1974.

O. H. Brownlee, University of Minnesota: Deputy Assistant Secretary, Tax Analysis Division, U.S. Treasury Department.

Burnham Campbell, University of Hawaii: visiting professor, Thammasat University, Bangkok, fall 1973.

David C. Campbell, University of Idaho: OAS resident expert in project evaluation, Ministry of Planning and Development, Trinidad and Tobago; Ministry of Finance and Planning, Barbados.

Barry R. Chiswick, Queens College and National Bureau of Economic Research: senior staff economist, Council of Economic Advisers, Aug. 1973.

Michael J. Farrell, California State University, Long Beach: Universidad de San Agustin, Arequipa, Peru, fall 1973.

Lehman B. Fletcher, Iowa State University: chief, Agricultural Economics and Sector Planning Division, Bureau of Technical Assistance, U.S. AID, Washington, D.C.

Leo V. Mayer, Iowa State University: senior staff economist, Council of Economic Advisers, to June 30, 1974.

Allan G. Pulsipher, Southern Illinois University: senior staff economist, Council of Economic Advisers, Aug. 1973–Aug. 1974.

James A. Seagraves, Iowa State University: water economist, Iowa-Peru project, Lima, Oct. 1973–Feb. 1975.

Jati K. Sengupta, Iowa State University: visiting professor, Indian Institute of Management, Calcutta, to Aug. 31, 1974.

Hylke Van de Wetering, Iowa State University: Chief-of-Party, Iowa-Peru project, Lima, Aug. 1973–75.

Resignations

Hang-Sheng Cheng, Iowa State University: Federal Reserve Bank of San Francisco.

Henry E. Cole, Tulane University: General Electric Co.

David Humphrey, Tulane University: California State University, San Francisco.

H. B. Howell, Iowa State University, June 30, 1973.

Philip H. White, University of British Columbia: Bronfman Organization in Europe, Aug. 1973.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, John W.) his present place of employment or enrollment: his new title (if any), his next place of employment (if known or if changed), and the date at which the change will occur.

C. Type each item on a separate 3×5 card, and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, November 1; *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1974 issue of the *Review* will carry the seventy-first list of doctoral dissertations in political economy in American universities and colleges. The list will specify doctoral degrees conferred during the academic year terminating June 1974. This announcement is an invitation to send us information for the preparation of the list. This announcement supercedes and replaces a letter which was sent annually from the managing editor's office.

The *Review* will publish in its December 1974 issue the names of those who will have been awarded the doctoral degree since June 1973 and the titles of their dissertations. Dissertation abstracts will no longer be published, as these are published elsewhere.

By June 30, please send us this information on 3×5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

JEL Classification No. ________

Name: LAST NAME IN CAPS: First Name, Initial

Institution Granting Degree: ________________

Degree Conferred (Ph.D. or D.B.A.) ________ Year ________

Dissertation Title: ________________

When degrees in economics are awarded under different names, such as Business Administration, Public Administration, or Industrial Relations, candidates in these fields whose training has been *primarily in economics* should be included.

All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

RICHARD T. ELY LECTURE

The Economics of Resources or the Resources of Economics

By Robert M. Solow*

It is easy to choose a subject for a distinguished lecture like this, before a large and critical audience with a wide range of interests. You need a topic that is absolutely contemporary, but somehow perennial. It should survey a broad field, without being superficial or vague. It should probably bear some relation to economic policy, but of course it must have some serious analytical foundations. It is nice if the topic has an important literature in the past of our subject—a literature which you can summarize brilliantly in about eleven minutes—but it better be something in which economists are interested today, and it should appropriately be a subject you have worked on yourself. The lecture should have some technical interest, because you can't waffle for a whole hour to a room full of professionals, but it is hardly the occasion to use a blackboard.

I said that it is easy to choose a subject for the Ely Lecture. It has to be, because twelve people, counting me, have done it.

I am going to begin with a quotation that could have come from yesterday's newspaper, or the most recent issue of the *American Economic Review.*

> Contemplation of the world's disappearing supplies of minerals, forests, and other exhaustible assets has led to demands for regulation of their exploitation. The feeling that these products are now too cheap for the good of future generations, that they are being selfishly exploited at too rapid a rate, and that in consequence of their excessive cheapness they are being produced and consumed wastefully has given rise to the conservation movement.

The author of those sentences is not Dennis Meadows and associates, not Ralph Nader and associates, not the President of the Sierra Club; it is a very eminent economic theorist, a Distinguished Fellow of this Association, Harold Hotelling, who died at the age of seventy-eight, just a few days ago. Like all economic theorists, I am much in his debt, and I would be happy to have this lecture stand as a tribute to him. These sentences appeared at the beginning of his article "The Economics of Exhaustible Resources," not in the most recent *Review*, but in the *Journal of Political Economy* for April 1931. So I think I have found something that is both contemporary and perennial. The world has been exhausting its exhaustible resources since the first cave-man chipped a flint, and I imagine the process will go on for a long, long time.

Mr. Dooley noticed that "th' Supreme Coort follows the iliction returns." He would be glad to know that economic theorists read the newspapers. About a year ago, having seen several of those respectable committee reports on the advancing scarcity of materials in the United States and the world, and having, like everyone else, been suckered into reading the *Limits to Growth*, I decided I ought to find out what economic theory has to say about the problems connected with exhaustible resources. I read some of the literature, including Hotelling's classic article—the theoretical literature on exhaustible resources is, fortunately, not very large—and began doing some work of my own on the problem of optimal social management of a stock of a nonrenewable but essential resource. I will be mentioning some of the results later. About the time I finished a first draft of my own paper and was patting myself on the back for having been clever enough to realize that there was in fact something still to be said on this important, contemporary but somehow perennial topic—just about then it seemed that every time the mail came it contained another paper by another economic theorist on the economics of exhaustible resources.[1] It was a little like trotting down to the sea, minding your own business like any nice independent rat, and then looking around and suddenly discovering that you're a lemming. Anyhow, I now have a nice collection of papers on the theory of exhaustible resources; and most of them are still unpublished, which is just the advantage I need over the rest of you.

---

* Professor of economics, Massachusetts Institute of Technology.

A pool of oil or vein of iron or deposit of copper in the ground is a capital asset to society and to its owner (in the kind of society in which such things have private owners) much like a printing press or a building or any other reproducible capital asset. The only difference is that the natural resource is not reproducible, so the size of the existing stock can never increase through time. It can only decrease (or, if none is mined for a while, stay the same). This is true even of recyclable materials; the laws of thermodynamics and life guarantee that we will never recover a whole pound of secondary copper from a pound of primary copper in use, or a whole pound of tertiary copper from a pound of secondary copper in use. There is leakage at every round; and a formula just like the ordinary multiplier formula tells us how much copper use can be built on the world's initial endowment of copper, in terms of the recycling or recovery ratio. There is always less ultimate copper use left than there was last year, less by the amount dissipated beyond recovery during the year. So copper remains an exhaustible resource, despite the possibility of partial recycling.

A resource deposit draws its market value, ultimately, from the prospect of extraction and sale. In the meanwhile, its owner, like the owner of every capital asset, is asking: What have you done for me lately? The only way that a resource deposit in the ground and left in the ground can produce a current return for its owner is by appreciating in value. Asset markets can be in equilibrium only when all assets in a given risk class earn the same rate of return, partly as current dividend and partly as capital gain. The common rate of return is the interest rate for that risk class. Since resource deposits have the peculiar property that they yield no dividend so long as they stay in the ground, in equilibrium the value of a resource deposit must be growing at a rate equal to the rate of interest. Since the value of a deposit is also the present value of future sales from it, after deduction of extraction costs, resource owners must expect the net price of the ore to be increasing exponentially at a rate equal to the rate of interest. If the mining industry is competitive, net price stands for market price minus marginal extraction cost for a ton of ore. If the industry operates under constant costs, that is just market price net of unit extraction costs, or the profit margin. If the industry is more or less monopolistic, as is frequently the case in extractive industry, it is the marginal profit—marginal revenue less marginal cost—that has to be growing, and expected to grow, proportionally like the rate of interest.

---

[1] *The Review of Economic Studies* will publish a group of them in the summer of 1974, including my own paper and others by Partha Dasgupta and Geoffrey Heal, Michael Weinstein and Richard Zeckhauser, and Joseph Stiglitz, from all of which I have learned a lot about this subject. I would especially like to thank Zeckhauser for conversation and correspondence, and for the kind of reading of the first draft of this Lecture that one only dares to hope to get because it is so close to Christmas. The final version reflects his comments.

This is the fundamental principle of the economics of exhaustible resources. It was the basis of Hotelling's classic article. I have deduced it as a condition of stock equilibrium in the asset market. Hotelling thought of it mainly as a condition of flow equilibrium in the market for ore: if net price is increasing like compound interest, owners of operating mines will be indifferent at the margin between extracting and holding at every instant of time. So one can imagine production just equal to demand at the current price, and the ore market clears. No other time profile for prices can elicit positive production in every period of time.

It is hard to overemphasize the importance of this tilt in the time profile for net price. If the net price were to rise too slowly, production would be pushed nearer in time and the resource would be exhausted quickly, precisely because no one would wish to hold resources in the ground and earn less than the going rate of return. If the net price were to rise too fast, resource deposits would be an excellent way to hold wealth, and owners would delay production while they enjoyed supernormal capital gains.

According to the fundamental principle, if we observe the market for an exhaustible resource near equilibrium, we should see the net price—or marginal profit—rising exponentially. That is not quite the same thing as seeing the market price to users of the resource rising exponentially. The price to consumers is the net price plus extraction costs, or the obvious analogy for monopoly. The market price can fall or stay constant while the net price is rising if extraction costs are falling through time, and if the net price or scarcity rent is not too large a proportion of the market price. That is presumably what has been happening in the market for most exhaustible resources in the past. (It is odd that there are not some econometric studies designed to find out just this. Maybe econometricians don't follow the iliction returns.) Eventually, as the extraction cost falls and the net price rises, the scarcity rent must come to dominate the movement of market price, so the market price will eventually rise, although that may take a very long time to happen. Whatever the pattern, the market price and the rate of extraction are connected by the demand curve for the resource. So, ultimately, when the market price rises, the current rate of production must fall along the demand curve. Sooner or later, the market price will get high enough to choke off the demand entirely. At that moment production falls to zero. If flows and stocks have been beautifully coordinated through the operations of futures markets or a planning board, the last ton produced will also be the last ton in the ground. The resource will be exhausted at the instant that it has priced itself out of the market. The Age of Oil or Zinc or Whatever It Is will have come to an end. (There is a limiting case, of course, in which demand goes asymptotically to zero as the price rises to infinity, and the resource is exhausted only asymptotically. But it is neither believable nor important.)

Now let us do an exercise with this apparatus. Suppose there are two sources of the same ore, one high-cost and the other low-cost. The cost difference may reflect geographical accessibility and transportation costs, or some geological or chemical difference that makes extraction cheap at one site and dear at the other. The important thing is that there are cost differences, though the final mineral product is identical from both sources.

It is easy to see that production from both sources cannot coexist in the market for any interval of time. For both sources to produce, net price for each of them must be growing like compound interest at the market rate. But they must market their ore at the same price, because the product is identical. That is arithmetically impossible, if their extraction costs differ.

So the story has to go like this. First one source operates and supplies the whole market. Its net price rises exponentially, and the market price moves correspondingly. At a certain moment, the first source is exhausted. At just that moment and not before, it must become economical for the second source to come into production. From then on, the world is in the single-source situation: the net price calculated with current extraction costs must rise exponentially until all production is choked off and the second source is exhausted. (If there are many sources, you can see how it will work.)

Which source will be used first? Your instinct tells you that the low-cost deposit will be the first one worked, and your instinct is right. You can see why, in terms of the fundamental principle. At the beginning, if the high-cost producer is serving the market, the market price must cover high extraction costs plus a scarcity rent that is growing exponentially. The low-cost producer would refrain from undercutting the price and entering the market only if his capital gains justify holding off and entering the market later. But just the reverse will be true. Any price high enough to keep the high-cost producer in business will tempt the low-cost producer to sell ore while the selling is good and invest the proceeds in any asset paying the market rate of interest. So it must be that the low-cost producer is the first to enter. Price rises and output falls. Eventually, at precisely the moment when the low-cost supply is exhausted, the price has reached a level at which it pays the high-cost producer to enter. From then on, *his* net price rises exponentially and production continues to fall. When cumulative production has exhausted the high-cost deposit, the market price must be such as to choke the demand off to zero—or else just high enough to tempt a still higher-cost source into production. And so it goes. Apart from market processes, it is actually socially rational to use the lower-cost deposits before the higher-cost ones.

You can take this story even further, as William Nordhaus has done in connection with the energy industry. Suppose that, somewhere in the background, there is a technology capable of producing or substituting for a mineral resource at relatively high cost, but on an effectively inexhaustible resource base. Nordhaus calls this a "backstop technology." (The nearest we now have to such a thing is the breeder reactor using U^{238} as fuel. World reserves of U^{238} are thought to be enough to provide energy for over a million years at current rates of consumption. If that is not a backstop technology, it is at least a catcher who will not allow a lot of passed balls. For a better approximation, we must wait for controlled nuclear fusion or direct use of solar energy. The sun will not last forever, but it will last at least as long as we do, more or less by definition.) Since there is no scarcity rent to grow exponentially, the backstop technology can operate as soon as the market price rises enough to cover its extraction costs (including, of course, profit on the capital equipment involved in production). And as soon as that happens, the market price of the ore or its substitute stops rising. The "backstop technology" provides a ceiling for the market price of the natural resource.

The story in the early stages is as I have told it. In the beginning, the successive grades of the resource are mined. The last and highest-cost source gives out just when the market price has risen to the point where the backstop technology becomes competitive. During the earlier phases, one imagines that resource companies keep a careful eye on the prospective costs associated with the backstop technology. Any laboratory success or failure that changes those prospective costs has instantaneous effects on the capital value of existing resource deposits, and on the most profitable rate of current production. In actual fact, those future costs have to be regarded as uncertain. A correct theory of market behavior and a correct theory of optimal social policy will have to take account of technological uncertainty (and perhaps also uncertainty about the true size of mineral reserves).

Here is a mildly concrete illustration of these principles. There is now a workable technology for liquefying coal—that is, for producing synthetic crude oil from coal.[2] Nordhaus puts the extraction-and-preparation cost at the equivalent of seven or eight 1970 dollars per barrel of crude oil, including amortization and interest at 10 percent on the plant; I have heard higher and lower figures quoted. If coal were available in unlimited amounts, that would be all. But, of course, coal is a scarce resource, though more abundant than drillable petroleum, so a scarcity rent has to be added to that figure, and the rent has to be increasing like the rate of interest during the period when coal is being used for this purpose.

In the meanwhile, the extraction and production cost for this technology is large compared with the scarcity rent on the coal input, so the market price at which the liquefied-coal-synthetic-crude activity would now be economic is rising more slowly than the rate of interest. It may even fall if there are cost-reducing technological improvements; and that is not unlikely, given that research on coal has not been splashed as liberally with funds as research on nuclear energy. In any case, political shenanigans and monopoly profits aside, scarcity rents on oil form a larger fraction of the market price of oil, precisely because it is a lower cost fuel. The price of a barrel of oil should therefore be rising faster than the implicit price at which synthetic crude from coal could compete. One day those curves will intersect, and that day the synthetic-crude technology will replace the drilled-petroleum technology.

Even before that day, the possibility of coal liquefaction provides a kind of ceiling for the price of oil. I say "kind of" to remind you that coal-mining and moving capacity and synthetic-crude plant cannot be created overnight. One might hope that the ceiling might also limit the consuming world's vulnerability to political shenanigans and monopoly profits. I suppose it does in some ultimate sense, but one must not slide over the difficulties: for example, who would want to make a large investment in coal liquefaction or coal gasification in the knowledge that the current price of oil contains a large monopoly element that could be cut, at least temporarily, if something like a price war should develop?

The fundamental principle of the economics of exhaustible resources is, as I have said, simultaneously a condition of flow equilibrium in the market for the ore and of asset equilibrium in the market for deposits. When it holds, it says quite a lot about the probable pattern of exploitation of a resource. But there are more than the usual reasons for wondering whether the equilibrium conditions have any explanatory value. For instance, the flow market that has to be cleared is not just one market; it is the sequence of markets for resource products from now until the date of exhaustion. It is, in other words, a sequence of futures markets, perhaps a long sequence. If the futures markets actually existed, we could perhaps accept the notion that their equilibrium configuration is stable; that might not be true, but it is at least the sort of working hypothesis we frequently accept as a way of getting on with business. But there clearly is not a full set of futures markets; natural-resource markets work with a combination of myopic flow transactions and rather more farsighted asset transactions. It is legitimate to ask whether observed resource prices are to be interpreted as approximations to equilibrium prices, or whether the equilibrium is so unstable that momentary prices are not only a bad indicator of equilibrium relationships, but also a bad guide to resource allocation.

---

[2] As best one can tell at the moment, shale oil is a more likely successor to oil and natural gas than either gasified or liquefied coal. The relevant costs are bound to be uncertain until more research and development has been done. I tell the story in terms of liquefied coal only because it is more picturesque that way.

That turns out not to be an easy question to answer. Flow considerations and stock considerations work in opposite directions. The flow markets by themselves could easily be unstable; but the asset markets provide a corrective force. Let me try to explain why.

The flow equilibrium condition is that the net price grow like compound interest at the prevailing rate. Suppose net prices are expected by producers to be rising too slowly. Then resource deposits are a bad way to hold wealth. Mine owners will try to pull out; and if they think only in flow terms, the way to get out of the resource business is to increase current production and convert ore into money. If current production increases, for this or any other reason, the current price must move down along the demand curve. So initially pessimistic price expectations on the part of producers have led to more pressure on the current price. If expectations about future price changes are responsive to current events, the consequence can only be that pessimism is reinforced and deepened. The initial disequilibrium is worsened, not eliminated, by this chain of events. In other words, the market mechanism I have just described is unstable. Symmetrical reasoning leads to the conclusion that if prices are initially expected to be rising too fast, the withholding of supplies will lead to a speculative run-up of prices which is self-reinforcing. Depending on which way we start, initial disequilibrium is magnified, and production is tilted either toward excessive current dumping or toward speculative withholding of supply. (Still other assumptions are possible and may lead to qualitatively different results. For instance, one could imagine that expectations focus on the price level rather than its rate of change. There is much more work to be done on this question.)

Such things have happened in resource markets; but they do not seem always to be happening. I think that this story of instability in spot markets needs amendment; it is implausible because it leaves the asset market entirely out of account. The longer run prospect is not allowed to have any influence on current happenings. Suppose that producers do have some notion that the resource they own has a value anchored somewhere in the future, a value determined by technological and demand considerations, not by pure and simple speculation. Then if prices are now rising toward that rendezvous at too slow a rate, that is indeed evidence that owning resource deposits is bad business. But that will lead not to wholesale dumping of current production, but to capital losses on existing stocks. When existing stocks have been written down in value, the net price can rise toward its future rendezvous at more or less the right rate. As well as being destabilized by flow reactions, the market can be stabilized by capitalization reactions. In fact the two stories can be made to merge: the reduction in flow price coming from increased current production can be read as a signal and capitalized into losses on asset values, after which near-equilibrium is reestablished.

I think the correct conclusion to be drawn from this discussion is not that either of the stories is more likely to be true. It is more complex: that in tranquil conditions, resource markets are likely to track their equilibrium paths moderately well, or at least not likely to rush away from them. But resource markets may be rather vulnerable to surprises. They may respond to shocks about the volume of reserves, or about competition from new materials, or about the costs of competing technologies, or even about near-term political events, by drastic movements of current price and production. It may be quite a while before the transvaluation of values—I never thought I could quote Nietzsche in an economics paper—settles down under the control of sober future prospects. In between, it may be a cold winter.

So far, I have discussed the economic theory of exhaustible resources as a partial-equilibrium market theory. The interest rate that more or less controls the whole process was taken as given to the mining industry by the rest of the economy. So was the demand curve for the resource itself. And when the market price of the resource has ridden up the demand curve to the point where the quantity demanded falls to zero, the theory says that the resource in question will have been exhausted.

There is clearly a more cosmic aspect to the question than this; and I do not mean to suggest that it is unimportant, just because it is cosmic. In particular, there remains an important question about the social interest in the pace of exploitation of the world's endowment of exhaustible natural resources. This aspect has been brought to a head recently, as everyone knows, by the various Doomsday forecasts that combine a positive finding that the world is already close to irreversible collapse from shortage of natural resources and other causes with the normative judgment that civilization is much too young to die. I do not intend to discuss those forecasts and judgments now—this convention already has one session devoted to just that—but I do want to talk about the economic issues of principle involved.

First, there is a proposition that will be second nature to everyone in this room. What I have called the fundamental principle of the economics of exhaustible resources is, among other things, a condition of competitive equilibrium in the sequence of futures markets for deliveries of the natural resource. This sequence extends out to infinity, even if the competitive equilibrium calls for the resource to be exhausted in finite time. Beyond the time of exhaustion there is also equilibrium: supply equals demand equals zero at a price simultaneously so high that demand is choked off and so low that it is worth no one's while to lose interest by holding some of the resource that long. Like any other competitive equilibrium with the right background assumptions, this one has some optimality properties. In particular, as Hotelling pointed out, the competitive equilibrium maximizes the sum of the discounted consumer-plus-producer surpluses from the natural resource, *provided* that society wishes to discount future consumer surpluses at the same rate that mine owners choose to discount their own future profits.

Hotelling was not so naive as to leap from this conclusion to the belief that *laissez-faire* would be an adequate policy for the resource industries. He pointed to several ways in which the background assumptions might be expected to fail: the presence of externalities when several owners can exploit the same underground pool of gas or oil; the considerable uncertainty surrounding the process of exploration with the consequent likelihood of wasteful rushes to stake claims and exploit, and the creation of socially useless windfall profits; and, finally, the existence of large monopolistic or oligopolistic firms in the extractive industries.

There is an amusing sidelight here. It is not hard to show that, generally speaking, a monopolist will exhaust a mine more slowly than a competitive industry facing the same demand curve would do. (Hotelling did not explore this point in detail, though he clearly knew it. He did mention the possibility of an extreme case in which competition will exhaust a resource in finite time and a monopolist only asymptotically.) The amusing thing is that if a conservationist is someone who would like to see resources conserved *beyond* the pace that competition would adopt, then the monopolist is the conservationist's friend. No doubt they would both be surprised to know it.

Hotelling mentions, but rather pooh-poohs, the notion that market rates of interest might exceed the rate at which society would wish to discount future utilities or consumer surpluses. I think a modern economist would take that possibility more seriously. It is certainly a potentially important question, because the discount rate determines the whole tilt of the equilibrium production schedule. If it is true that the market rate of interest exceeds the social rate of time preference, then scarcity rents and market prices will rise faster than they "ought to" and production will have to fall correspondingly faster along the demand curve. Thus the resource will be exploited too fast and exhausted too soon.

The literature has several reasons for expecting that private discount rates might be systematically higher than the correct social rate of discount. They fall into two classes. The first class takes it more or less for granted that society ought to discount utility and consumption at the same rates as reflective individuals would discount their own future utility and consumption. This line of thought then goes on to suggest that there are reasons why this might not happen. One standard example is the fact that individuals can be expected to discount for the riskiness of the future, and some of the risks for which they will discount are not risks to society but merely the danger of transfers within the society. Since there is not a complete enough set of insurance markets to permit all these risks to be spread properly, market interest rates will be too high. Insecurity of tenure, as William Vickrey has pointed out, is a special form of uncertainty with particular relevance to natural resources.

A second standard example is the existence of various taxes on income from capital; since individuals care about the after-tax return on investment and society about the before-tax return, if investment is carried to the point where the after-tax yield is properly related to the rate of time preference, the before-tax profitability of investment will be too high. I have nothing to add to this discussion.

The other class of reasons for expecting that private discount rates are too high and will thus distort intertemporal decisions away from social optimality denies that private time preference is the right basis for intertemporal decisions. Frank Ramsey, for instance, argued that it was ethically indefensible for society to discount future utilities. Individuals might do so, either because they lack imagination (Böhm-Bawerk's "defective telescopic faculty") or because they are all too conscious that life is short. In social decision-making, however, there is no excuse for treating generations unequally, and the time-horizon is, or should be, very long. In solemn conclave assembled, so to speak, we ought to act as if the social rate of time preference were zero (though we would simultaneously discount future *consumption* if we expect the future to be richer than the present). I confess I find that reasoning persuasive, and it provides another reason for expecting that the market will exhaust resources too fast.

This point need not be divorced so completely from individual time preference. If the whole infinite sequence of futures markets for resource products could actually take place and find equilibrium, I might be inclined to accept the result (though I would like to know who decides the initial endowments within and between generations). But of course they cannot take place. There is no way to collect bids and offers from everyone who will ever live. In the markets that actually do take place, future generations are represented only by us, their eventual ancestors. Now generations overlap, so that I worry about my children, and they about theirs, and so on. But it does seem fundamentally implausible that there should be anything *ex post* right about the weight that is actually given to the welfare of those who will not live for another thousand years. We have actually done quite well at the hands of *our* ancestors. Given how poor they were and how rich we are, they might properly have saved less and consumed more. No doubt they never expected the rise in income per head that has made us so much richer than they ever dreamed was possible. But that only reinforces the point that the future may be too important to be left to the accident of mistaken expectations and the ups and downs of the Protestant ethic.

Several writers have studied directly the problem of defining and characterizing a socially-optimal path for the exploitation of a given pool of exhaustible resources. The idea is familiar enough: instead of worrying about market responses, one imagines an idealized planned economy, constrained only by its initial endowment, the size of the labor force, the available technology, and the laws of arithmetic. The planning board then has to find the best feasible development for the economy. To do so, it needs a precise criterion for comparing different paths, and that is where the social rate of time preference plays a role.

It turns out that the choice of a rate of time preference is even more critical in this situation than it is in the older literature on optimal capital accumulation without any exhaustible resources. In that theory, the criterion usually adopted is the maximization of a discounted sum of one-period social welfare indicators, depending on consumption per head, and summed over all time from now to the infinite future. The typical result, depending somewhat on the particular assumptions made, is that consumption per head rises through time to a constant plateau defined by the "modified Golden Rule." In that ultimate steady state, consumption per head is lower the higher is the social rate of discount; and, correspondingly, the path to the steady state is characterized by less saving and more interim consumption, the higher the social rate of discount. That is as it should be: the main beneficiaries of a high level of ultimate steady-state consumption are the inhabitants of the distant future, and so, if the planning board discounts the future very strongly, it will choose a path that favors the near future over the distant future.

When one adds exhaustible resources to the picture, the social rate of time preference can play a similar, but even more critical, role. As a paper by Geoffrey Heal and Partha Dasgupta and one of my own show, it is possible that the optimal path with a positive discount rate should lead to consumption per head going asymptotically to zero,-whereas a zero discount rate leads to perpetually rising consumption per head. In other words, even when the technology and the resource base could permit a plateau level of consumption per head, or even a rising standard of living, positive social time preference might in effect lead society to prefer eventual extinction, given the drag exercised by exhaustible resources. Of course, it is part of the point that it is the planning board in the present that plans for future extinction: nobody has asked the about-to-become-defunct last generation whether *it* approved of weighting its satisfactions less than those of its ancestors.

Good theory is usually trying to tell you something, even if it is not the literal truth. In this context, it is not hard to interpret the general tenor of the theoretical indications. We know in general that even well-functioning competitive markets may fail to allocate resources properly over time. The reason, I have suggested, is because, in the nature of the case, the future brings no endowment of its own to whatever markets actually exist. The intergenerational distribution of income or welfare depends on the provision that each generation makes for its successors. The choice of a social discount rate is, in effect, a policy decision about that intergenerational distribution. What happens in the planning parable depends very much—perhaps dramatically—on that choice; and one's evaluation of what happens in the market parable depends very much on whether private choices are made with a discount rate much larger than the one a deliberate policy decision would select. The pure theory of exhaustible resources is trying to tell us that, if exhaustible resources really matter, then the balance between present and future is more delicate than we are accustomed to think; and then the choice of a discount rate can be pretty important and one ought not to be too casual about it.

In my own work on this question, I have sometimes used a rather special criterion that embodies sharp assumptions about intergenerational equity: I have imposed the requirement that consumption per head be constant through time, so that no generation is favored over any other, and asked for the largest steady consumption per head that can be maintained forever, given all the constraints including the finiteness of resources. This criterion, like any other, has its pluses and its minuses and I am not committed to it by any means. Like the standard criterion—the discounted sum of one-period utilities—this one will always pick out an *efficient* path, so one at least gets the efficiency conditions out of the analysis. The highest-constant-consumption criterion also has the advantage of highlighting the crucial importance of certain technological assumptions.

It is clear without any technical apparatus that the seriousness of the resource-exhaustion problem must depend in an important way on two aspects of the technology: first, the likelihood of technical progress, especially natural-resource-saving technical progress, and, second, the ease with which other factors of production, especially labor and reproducible capital, can be substituted for exhaustible resources in production.

My own practice, in working on this problem, has been to treat as the central case (though not the only case) the assumption of zero technological progress. This is not because I think resource-saving inventions are unlikely or that their capacity to save resources is fundamentally limited. Quite the contrary—if the future is anything like the past, there will be prolonged and substantial reductions in natural-resource requirements per unit of real output. It is true, as pessimists say, that it is just an assumption and one cannot be sure; but to assume the contrary is also an assumption, and a much less plausible one. I think there is virtue in analyzing the zero-technical-progress case because it is easy to see how technical progress can relieve and perhaps eliminate the drag on economic welfare exercised by natural-resource scarcity. The more important task for theory is to try to understand what happens or can happen in the opposite case.

As you would expect, the degree of substitutability is also a key factor. If it is very easy to substitute other factors for natural resources, then there is in principle no "problem." The world can, in effect, get along without natural resources, so exhaustion is just an event, not a catastrophe. Nordhaus's notion of a "backstop technology" is just a dramatic way of putting this case; at some finite cost, production can be freed of dependence on exhaustible resources altogether.

If, on the other hand, real output per unit of resources is effectively bounded—cannot exceed some upper limit of productivity which is in turn not too far from where we are now—then catastrophe is unavoidable. In-between there is a wide range of cases in which the problem is real, interesting, and not foreclosed. Fortunately, what little evidence there is suggests that there is quite a lot of substitutability between exhaustible resources and renewable or reproducible resources, though this is an empirical question that could absorb a lot more work than it has had so far.

Perhaps the most dramatic way to illustrate the importance of substitutability, and its connection with Doomsday, is in terms of the permanent sustainability of a constant level of consumption. In the simplest, most aggregative, model of a resource-using economy one can prove something like the following: if the elasticity of substitution between exhaustible resources and other inputs is unity or bigger, and if the elasticity of output with respect to reproducible capital exceeds the elasticity of output with respect to natural resources, then a constant population can maintain a positive constant level of consumption per head forever. This permanently maintainable standard of living is an increasing, concave, and unbounded function of the initial stock of capital. So the drag of a given resource pool can be overcome *to any extent* if only the initial stock of capital is large enough. On the other hand, if the elasticity of substitution between natural resources and other inputs is less than one, or if the elasticity of output with respect to resources exceeds the elasticity of output with respect to reproducible capital, then the largest constant level of consumption sustainable forever with constant population is—zero. We know much too little about which side of that boundary the world is on—technological progress aside—but at least the few entrails that have been read seem favorable.[3]

Perhaps I should mention that when I say "forever" in this connection, I mean "for a very long time." The mathematical reasoning does deal with infinite histories, but actually life in the solar system will only last for a finite time, though a very long finite time, much longer than this lecture, for instance. That is why I think it takes economics as well as the entropy law to answer our question.

I began this lecture by talking of the conditions for competitive equilibrium in the market for natural resources. Now I have been talking of centralized planning optima. As you would expect, it turns out that under the standard assumptions, the Hotelling rule, the fundamental principle of natural-resource economics, is a necessary condition for efficiency and therefore for social optimality. So there is at least a prayer that a market-guided system might manage fairly well. But more than the Hotelling condition is needed.

---

[3] See pp. 60–70 in William D. Nordhaus and James Tobin.

I have already mentioned one of the extra requirements for the intertemporal optimality of market allocations: it is that the market discount future profits at the same rate as the society would wish to discount the welfare of future inhabitants of the planet. This condition is often given as an argument for public intervention in resource allocation because—as I have also mentioned—there are reasons to expect market interest rates to exceed the social rate of time preference, or at least what philosophers like us think it ought to be. If the analysis is right, then the market will tend to consume exhaustible resources too fast, and corrective public intervention should be aimed at slowing down and stretching out the exploitation of the resource pool. There are several ways that could be done, in principle, through conservation subsidies or a system of graduated severance taxes, falling through time.

Realistically speaking, however, when we say "public intervention" we mean rough and ready political action. An only moderately cynical observer will see a problem here: it is far from clear that the political process can be relied on to be more future-oriented than your average corporation. The conventional pay-out period for business is of the same order of magnitude as the time to the next election, and transferring a given individual from the industrial to the government bureaucracy does not transform him into a guardian of the far future's interests. I have no ready solution to this problem. At a minimum, it suggests that one ought to be as suspicious of uncritical centralization as of uncritical free-marketeering. Maybe the safest course is to favor specific policies—like graduated severance taxes—rather than blanket institutional solutions.

There is another, more subtle, extra requirement for the optimality of the competitive market solution to the natural-resource problem. Many patterns of exploitation of the exhaustible-resource pool obey Hotelling's fundamental principle myopically, from moment to moment, but are wrong from a very long-run point of view. Such mistaken paths may even stay very near the right path for a long time, but eventually they veer off and become bizarre in one way or another. If a market-guided system is to perform well over the long haul, it must be more than myopic. Someone—it could be the Department of the Interior, or the mining companies, or their major customers, or speculators—must always be taking the long view. They must somehow notice in advance that the resource economy is moving along a path that is bound to end in disequilibrium of some extreme kind. If they do notice it, and take defensive actions, they will help steer the economy from the wrong path toward the right one.[4] Usually the "wrong" path is one that leads to exhaustion at a date either too late or too soon; anyone who perceives this will be motivated to arbitrage between present and future in ways that will push the current price toward the "right" path.[5]

It is interesting that this need for someone to take the long view emerged also when the question at hand was the potential instability of the market for natural resources if it concentrates too heavily on spot or flow decisions, and not enough on future or stock decisions. In that context too, a reasonably accurate view of the long-term prospects turns out to be a useful, maybe indispensable, thing for the resource market to have.

---

[4] This sort of process has been studied in a different context by Frank Hahn and by Karl Shell and Joseph Stiglitz.

[5] For example, suppose the current price is too low, in the sense that, if it rises according to the current principle, the demand path will be enough to exhaust the resource before the price has risen high enough to choke demand to zero. A clever speculator would see that there will be money to be made just after the date of exhaustion, because anyone with a bit of the resource to sell could make a discrete jump in the price and still find buyers. Such a speculator would wish to buy now

This lecture has been—as Kenneth Burke once said about the novel—words, all words. Nevertheless, it has been a discourse on economic theory, not on current policy. If some of you have been daydreaming about oil and the coming winter, I assure you that I have been thinking about shadow prices and transversality conditions at infinity. If I turn briefly to policy at the end, it is not with concrete current problems in mind. After all, nothing I have been able to say takes account of the international oil cartel, the political and economic ambitions of the Middle Eastern potentates, the speeds of adjustment to surprises in the supply of oil, or the doings of our own friendly local oligopolists. The only remarks I feel entitled to make are about the long-run pursuit of a general policy toward exhaustible resources.

Many discussions of economic policy—macroeconomics aside—boil down to a tension between market allocation and public intervention. Marketeers keep thinking about the doughnut of allocative efficiency and informational economy and *dirigistes* are impressed with the size of the hole containing externalities, imperfections, and distributional issues. So it is with exhaustible resources. One is impressed with what a system of ideal markets, including futures markets, can accomplish in this complicated situation; and one can hardly miss seeing that our actual oligopolistic, politically involved, pollution-producing industry is not exactly what the textbook ordered. I have nothing new to add to all that. The unusual factor that the theory of exhaustible resources brings to the fore is the importance of the long view, and the value of reasonable information about reserves, technology, and demand in the fairly far future.

This being so, one is led to wonder whether public policy can contribute to stability and efficiency along those lines. One possibility is the encouragement of organized futures trading in natural resource products. To be useful, futures contracts would have to be much longer-term than is usual in the futures markets that now exist, mostly for agricultural products. I simply do not know enough to have an opinion about the feasibility of large scale futures trading, or about the ultimate contribution that such a reform would make to the stability and efficiency of the market for resource products. But in principle it would seem to be a good idea.

The same considerations suggest that the market for exhaustible resources might be one of the places in the economy where some sort of organized indicative planning could play a constructive role. This is not an endorsement of centralized decision-making, which is likely to have imperfections and externalities of its own. Indeed it might be enough to have the government engaged in a continuous program of information-gathering and dissemination covering trends in technology, reserves and demand. One could at least hope to have professional standards govern such an exercise. I take it that the underlying logic of indicative planning is that some comparison and coordination of the main participants in the market, including the

---

and hold for sale then. But that action would tend to raise the current price (and, by the fundamental principle, the whole price path) and reduce demand, so that the life of the resource would be prolonged. The speculation is thus corrective.

government, could eliminate major errors and resolve much uncertainty. In the case of exhaustible resources, it could have the additional purpose of generating a set of consistent expectations about the distant future. In this effort, the pooling of information and intentions from both sides of the market could be useful, with the effect of inducing behavior that will steer the economy away from ultimately inferior paths. It is also likely, as Adam Smith would have warned, that a certain amount of conspiracy against the public interest might occur in such sessions, so perhaps they ought to be recorded and the tapes turned over to Judge Sirica, who will know what to do with them.

REFERENCES

P. Dasgupta and G. Heal, "The Optimal Depletion of Exhaustible Resources," *Rev. Econ. Stud.*, forthcoming, 1974.

F. H. Hahn, "Equilibrium Dynamics with Heterogeneous Capital Goods," *Quart. J. Econ.*, Nov. 1966, *80*, 633–646.

H. Hotelling, "The Economics of Exhaustible Resources," *J. Polit. Econ.*, April 1931, *39*, 137–175.

W. D. Nordhaus, "The Allocation of Energy Resources," *Brookings Papers on Econ. Activ.*, forthcoming.

——— and J. Tobin, "Is Economic Growth Obsolete?" in National Bureau of Economic Research, *Economic Growth*, 50th Anniversary Colloq. V, New York 1972.

K. Shell and J. E. Stiglitz, "The Allocation of Investment in a Dynamic Economy," *Quart. J. Econ.*, Nov. 1967, *81*.

R. M. Solow, "Intergenerational Equity and Exhaustible Resources," *Rev. Econ. Stud.*, forthcoming, 1974.

J. E. Stiglitz, "Growth with Exhaustible Natural Resources," *Rev. Econ. Stud.*, forthcoming, 1974.

M. Weinstein and R. Zeckhauser, "Use Patterns for Depletable and Recyclable Resources," *Rev. Econ. Stud.*, forthcoming, 1974.

THE WORLD DYNAMICS OF ECONOMIC GROWTH

The Economics of the Steady State

By Herman E. Daly*

> But if your theory is found to be against the second law of thermodynamics, I can give you no hope; there is nothing for it but to collapse in deepest humiliation.
>
> *Sir Arthur Eddington*

My title is somewhat pretentious since at present this "new economics" consists only of a definition of a steady-state economy, some arguments for its necessity and desirability, and some disciplined speculations on its appropriate institutions and the problem of transition, each of which will be briefly discussed below.

I. What is a Steady-State Economy?

A steady-state economy is defined by constant stocks of physical wealth (artifacts) and a constant population, each maintained at some chosen, desirable level by a low rate of throughput—i.e., by low birth rates equal to low death rates and by low physical production rates equal to low physical depreciation rates, so that longevity of people and durability of physical stocks are high. The throughput flow, viewed as the cost of maintaining the stocks, begins with the extraction (depletion) of low entropy resources at the input end, and terminates with an equal quantity of high entropy waste (pollution) at the output end. The throughput is the inevitable cost of maintaining the stocks of people and artifacts and should be minimized subject to the maintenance of a chosen level of stocks (Kenneth E. Boulding).

The services (want satisfaction) yielded by the stocks of artifacts (and people) are the ultimate benefit of economic activity, and the throughput is the ultimate cost. The stock of physical wealth is an accumulated flow of throughput, and thus in the final analysis is a cost. Ultimate efficiency is the ratio of service to throughput. But to yield a service, the throughput flow must be first accumulated into stocks even if of short duration. It is the existence of a table or a doctor at a point in time that yields services, not their gradual depreciation nor the productive process by which they are replaced. Stocks are intermediate magnitudes that yield services and require throughput for maintenance and replacement. This may be expressed in the equation:

$$(1)\quad \frac{\text{Ultimate}}{\text{Efficiency}} = \overset{(1)}{\frac{\text{Service}}{\text{Throughput}}} = \overset{(2)}{\frac{\text{Service}}{\text{Stock}}} \times \overset{(3)}{\frac{\text{Stock}}{\text{Throughput}}}$$

Since by definition stocks are constant at a level corresponding to some concept of sufficiency or maturity, progress in the steady state consists in increasing ultimate efficiency (ratio 1) in two ways: by maintaining the stock with less throughput (increase ratio 3 or "maintenance efficiency") and by getting more service per unit of time from the same stock (increase ratio 2 or "service efficiency"). The laws of thermodynamics provide a theoretical limit to the improvement of maintenance efficiency. Whether there is any theoretical limit to increase in service efficiency resulting from the limits of the human stomach and nervous system is less clear, but in my opinion likely.

---

* Professor of economics, Louisiana State University.

Over short periods of time the throughput cost of maintaining the constant stock may decrease due to improvements in maintenance efficiency, but over the long run it must increase because as better grade (lower entropy) sources of raw materials are used up, it will be necessary to process ever larger amounts of materials using ever more energy and capital equipment to get the same quantity of needed mineral. Thus a steady-state economy, as here defined, does not imply constant throughput, much less static technology, nor does it imply eternal life for the economic system. It is simply a strategy for good stewardship, for maintaining our spaceship and permitting it to die of old age rather than from the cancer of growthmania. It is basically an extension of the demographers' model of a stationary population to include the populations of physical artifacts, and the fundamental idea is found in John Stuart Mill's discussion of the stationary state of classical economics.

The term "economic growth" conventionally refers to an increase in the flow of "real *GNP*," which is a value *index* of the *physical* flow of throughput. The (measurable) throughput is, in turn, an index of (unmeasurable) service only if ratios 2 and 3 (or their product) are constant or increasing. This may have been the case in the past, but for the future it is doubtful. As the growing throughput pushes against biophysical limits, it provokes a decline in service efficiency (more of the stock must be devoted to the defensive use of repairing life support systems that formerly provided free services). Also, since our institutions are geared to a continually increasing throughput, we may willingly lower maintenance efficiency for the sake of permitting a larger throughput (e.g. planned obsolescence and fashion). If someone wants to redefine "economic growth" as an increase in nonmaterial services and then argue that it can and should grow forever, he is free to do so. But this hardly constitutes a refutation of the steady-state economy, which is defined in terms of measurable physical stocks, not unmeasurable psychic fluxes.

Nor are the levels at which the stocks of people and artifacts are maintained necessarily frozen for all eternity. As a result of technical and moral evolution it may become both possible and desirable to grow or to decline to a different level. But then growth or decline would be seen as a temporary transition from one steady state to another and not as the norm for a healthy economy. Technical and moral change would lead growth rather than being blindly pushed down the path of least resistance by the growth juggernaut.

At what point should growth in stocks and maximization of production flow give way to stock maintenance and the minimization of the production flow? There are a large number of steady-state levels of stocks to choose from, and such a choice is a difficult problem of ecology and ethics. But our inability to define the optimum level does not mean that we will not someday discover that we have grown beyond it. It is more important to learn to be stable at existing or nearby levels than to know in advance which level is optimal. Knowledge of the latter without the former merely allows us to recognize and wave goodbye to the optimum as we grow through it. Besides, the optimum may well be a broad plateau within which one place is as good as another as long as we don't go too near the edge.

The radical change implied by a steady state is evident from the foregoing, and from W. W. Rostow's characterization of our present economy of high mass consumption, to which all countries unrealistically aspire, as one "in which compound interest becomes built, as it were, into our habits and institutions" (p. 7). This built-in exponential growth and its unfortunate consequences constitute the theme of the much-maligned little book, *Limits to Growth*, by D. H. Meadows et al. Before discussing the radical departure of "deinstitutionalizing" compound interest or at least uncoupling it from all physical dimensions, we must consider whether such a change is really necessary and/or desirable.

II. The Necessity and Desirability of the Steady State

Our economy is a subsystem of the earth, and the earth is apparently a steady-state open system. The subsystem cannot grow beyond the frontiers of the total system and, if it is not to disrupt the functioning of the latter, must at some much earlier point conform to the steady-state mode. The technocratic project of redesigning the world (substituting technosphere for ecosphere) so as to allow for indefinite economic growth is a bit of hubris that has received the insufficiently pejorative label of "growthmania."

The conceptual roots of growthmania are to be found in the orthodox doctrines of "relative scarcity" and "absolute wants." Relative (or "Ricardian") scarcity refers to the scarcity of a particular resource relative to another resource or to a lower quality of the same resource. Absolute (or "Malthusian") scarcity refers to the scarcity of all resources in general, relative to population and per capita consumption levels. The solution to relative scarcity is substitution. Absolute scarcity assumes that all economical substitutions are made so that the total burden of absolute scarcity is minimized but still exists and may still increase. Even an efficiently borne burden can become too heavy. Substitution is always of one form of low entropy matter-energy for another. There is no substitute for low entropy itself, and low entropy is scarce, both in its terrestrial source (finite stocks of concentrated fossil fuels and minerals) and in its solar source (a fixed rate of inflow of solar energy). (See Georgescu-Roegen.) Both the human economy and the nonhuman part of the biosphere depend on the same limited budget of low entropy and on the allocative pattern which that budget has evolved over millennia. The entropy of the human sector is reduced and kept low by the continual importation of low entropy from, and exportation of high entropy to, the nonhuman sector (Daly, 1968). If too much low entropy is diverted to economic growth in the human sector, or if too many evolutionary allocative patterns are disrupted in the process of diversion, then the complex life support systems of the biosphere will begin to fail. Growth in population and per capita consumption result in increasing absolute scarcity, which is manifested in the increasing prevalence of "external costs"—i.e., the system becomes more generally sensitive to particular interferences as the web of general interdependence is stretched ever tighter by growth in the populations of people and artifacts and the resulting stress on the entropy budget.

Orthodox economic theory has assumed that all scarcity is relative: "Nature imposes particular scarcities, not an inescapable general scarcity" (Harold J. Barnett and Chandler Morse, p. 11). Therefore the answer to scarcity is always substitution, and since relative price changes induce substitution, the policy recommendation is "internalization of externalities," usually via pollution taxes. The following statement is representative of orthodox complacency: " . . . the problem of environmental pollution is a simple matter of correcting a minor resource misallocation by means of pollution charges . . . " (Wilfred Beckerman, p. 327). But price rigging by itself is ineffective in coping with increasing absolute scarcity since its mode of operation is only to induce substitution. What substitute is there for resources in general, for low entropy? How is it possible to raise the *relative* price of *all* resources? Attempts to do so result in inflation rather than substitution.

A similar distinction between absolute and relative wants has also been obscured by orthodox economics. Following Keynes we may define absolute wants as those we feel independent of the situation of our fellow human beings. Relative wants are those that we feel only if their satisfaction makes us feel superior to our fellows. The importance of this distinction is that only relative wants are infinite and that relative wants cannot be universally satisfied by growth because the relative satisfactions of the elite are cancelled as growth raises the general level. This effect can be avoided, and often is, by allowing growth to increase inequality so that the relatively well off become relatively better off. But it is quite impossible for everyone to become better off relative to everyone else. In spite of this extremely important distinction, orthodox theory assumes that wants *in general* are insatiable and extends to all wants the dignity of absolute status—i.e., the satisfaction of relative and absolute wants is considered equally legitimate and equally capable of satisfaction in the aggregate by means of economic growth. The assumption of equal legitimacy is a value judgment (though it is treated by many economists as the avoidance of a value judgment), and the assumption of equal capability of satisfaction is either a logical error or an implicit acceptance of a value judgment in favor of increasing inequality.

The implication of the dogmas of the relativity of all scarcity and the absoluteness of all wants is growthmania. If there is no absolute scarcity to limit the *possibility* of growth (infinite substitutability of relatively abundant for relatively scarce resources) and no merely relative wants to limit the *desirability* or efficacy of growth (wants in general are infinite and all wants are equally worthy and capable of satisfaction by growth), then "growth forever and the more the better" is the logical consequence. It is also the *reductio ad absurdum* that exposes the growth orthodoxy as a rigorous exercise in wishful thinking, as a theory that is against the second law of thermodynamics as well as against common sense. It is simply a brute fact that there is such a thing as absolute scarcity and such a thing as relative wants. Furthermore, these latter categories eventually become dominant at the margin as growth continues. The implication of absolute scarcity and relative wants is the opposite of growthmania, namely, the steady state.

At this point the growthmaniacs usually make a burnt offering to the god of technology: surely economic growth can continue indefinitely because technology will continue to "grow exponentially" as it has in the past. This elaborately misses the point. The alleged "exponential growth" of technology is not directly measurable and is only inferred from the permissive role that it has played in making possible the measured exponential growth in the physical magnitudes of production, depletion and pollution (i.e., the throughput). Such technical progress is more a part of the problem than the solution. What must be appealed to is a *qualitative change* in the direction of technical progress, not a continuation of alleged quantitative trends. The institutions to be discussed in the next section seek to induce just such a change toward resource-saving technology and patterns of living, and to a greater reliance on solar energy and renewable resources. But we can be fairly certain that no new technology will abolish absolute scarcity because the laws of thermodynamics apply to all possible technologies. No one can be absolutely certain that we will not some day discover perpetual motion and how to create and destroy matter and energy. But the reasonable assumption for economists is that this is an unlikely prospect and that while technology will continue to pull rabbits out of hats, it will not pull an elephant out of a hat—much less an infinite series of ever-larger elephants!

But the ideology of growth continues to transcend the ordinary logic of elementary economics. Growth is the basis of national power and prestige. Growth offers the prospect of prosperity for all with sacrifice by none. It is a substitute for redistribution. The present sins of poverty and injustice will be washed away in a future sea of abundance, vouchsafed by the amazing grace of compound interest. This evasion, common to both capitalism and communism, was never totally honest. It is now increasingly exposed as absurd.

III. Speculations on the Steady State

The first design principle disciplining our speculations on institutions is to provide the necessary social control with a minimum sacrifice of personal freedom, to provide macro stability while allowing for micro variability, to combine the macro static with the micro dynamic. A second design principle, closely related to the first, is to maintain considerable slack between the actual environmental load and the maximum carrying capacity. The closer the actual approaches the maximum, the more rigorous, finely tuned, and micro oriented our controls will have to be. We lack the knowledge and ability to assume detailed central control of the spaceship, even if such were desirable, so therefore we should leave it on "automatic pilot" as it has been for eons. But the automatic pilot only works if the actual load is small relative to the maximum. A third design principle, important for making the transition, is to start from existing initial conditions rather than an imaginary "clean slate," and a fourth is to build in the ability to tighten constraints gradually. Minimum faith is placed in our ability to plan a detailed blueprint for a new society. Maximum faith is placed in the basic regenerative powers of life and in the possibility of moral growth, once the root physical process of degeneration (unlimited growth) is arrested.

The kinds of institutions required follow directly from the definition. We need (1) an institution for stabilizing population, (2) an institution for stabilizing physical wealth and keeping throughput below ecological limits, and, less obviously but most importantly, (3) an institution limiting the degree of inequality in the distribution of the constant stocks among the constant population since growth can no longer be appealed to as the answer to poverty.

What specific institutions can perform these functions and are most in harmony with the general design principles discussed above? Elsewhere I have outlined a model and can here only briefly describe it (Daly, 1973, 1974a). The model builds on the existing institutions of private property and the price system and is thus fundamentally conservative. But it extends these institutions to areas previously not included: control of aggregate births (marketable birth license plan as first proposed by Boulding) and control of aggregate depletion of basic resources (depletion quotas auctioned by the government). Extending the market, under the discipline of aggregate quotas, to these vital areas is necessary to deal with increasing absolute scarcity since, as argued above, price controls deal only with relative scarcity. Quantitative limits are set with reference to ecological and ethical criteria, and the price system is then allowed, by auction and exchange, to allocate depletion quotas and birth quotas efficiently. The throughput is controlled at its input (depletion) rather than at the pollution end because physical control is easier at the point of lower entropy. Orthodox economics suggests price controls at the output end (pollution taxes), while steady-state economics suggests quantitative controls at the input end (depletion quotas).

With more vital areas of life officially subject to the discipline of the price system, it will become more urgent to establish the institutional preconditions of free and mutually beneficial exchange, namely, to limit the degree of inequality in the distribution of income and wealth and to limit the monopoly power of corporations. A distributist institution establishing a minimum income and a maximum income and wealth would go a long way toward achieving that end, while leaving room for differential reward and incentives within reasonable limits. There might be one set of limits for individuals, one for families, and one for corporations. Natural monopolies should be publicly owned and operated.

Birth quotas, depletion quotas, and distributive limits can all be varied continuously and applied with any degree of gradualism desired. Moreover, all three control points are price system parameters, and altering them does not interfere with the static allocative efficiency of the market. Externalities involving ecological, demographic, and distributive issues are "externalized" by means of quotas rather than "internalized" in rigged market prices. Yet the effect is much the same in that prices rise to reflect previously unaccounted dimensions of scarcity, and prices become a safer guide to market decisions. The net advantage of the quota scheme is that it limits aggregate throughput, whereas price controls merely alter through-put composition, providing a useful fine-tuning supplement to quotas, but not a substitute. The higher resource prices resulting from limited depletion would have the dynamic effect of inducing resource-saving technology and a shift to greater dependence on solar energy and renewable resources. The receipts of the depletion quota auction could help finance the minimum income. The marketable birth license plan would also have an equalizing effect on per capita income distribution.

Such institutional change is obviously not on the political agenda for 1974. Nor should it be since it is speculative, has not had the benefit of widespread professional criticism, and thus may contain terrible mistakes. But mistakes will not be discovered and better ideas will not be offered unless economists awake from the dogmatic slumber of growthmania induced by the soporific doctrines of relative scarcity and absolute wants and put the steady-state paradigm on the agenda for academic debate.

REFERENCES

H. J. Barnett and C. Morse, *Scarcity and Growth*, Baltimore 1963.

W. Beckerman, "Economists, Scientists, and Environmental Catastrophe," *Oxford Econ. Papers*, Nov. 1972, *24*, 327.

F. H. Bormann and W. R. Burch, eds., *Growth, Limits and the Quality of Life*, San Francisco 1974 (forthcoming).

K. E. Boulding, *Economics as a Science*, New York 1970.

J. Culbertson, *Economic Development: An Ecological Approach*, New York 1971.

H. E. Daly, "On Economics as a Life Science," *J. Polit. Econ.*, May/June 1968, *76*, 392–406.

———, "In Defense of a Steady-State Economy," *Amer. J. Agri. Econ.*, Dec. 1972, *54*, 945–954.

———, ed., *Toward a Steady-State Economy,* San Francisco 1973.

———, "Long Run Environmental Constraints and Trade-Offs Between Human and Artifact Populations," *Internat. Populat. Conf.,* IUSSP, Liege 1973, *3,* 453–460.

———, "A Model for a Steady-State Economy" in Bormann and Burch, 1974a.

———, "Steady-State Economics Versus Growthmania: A Critique of the Orthodox Conceptions of Growth, Wants, Scarcity, and Efficiency," *Policy Scien.* (forthcoming) Summer 1974b.

Editors of *The Ecologist,* "A Blueprint for Survival," *The Ecolo.,* Jan. 1972.

N. Georgescu-Roegen, *The Entropy Law and the Economic Process,* Cambridge, Mass. 1971.

D. H. Meadows et al., *The Limits to Growth,* New York 1972.

D. L. Meadows and D. H. Meadows, eds., *Toward Global Equilibrium: Collected Papers,* Cambridge, Mass., 1973.

W. Ophuls, *Prologue to a Political Theory of the Steady State,* (unpublished) Ph.D. dissertation in political science, Yale University, 1973.

W. W. Rostow, *The Stages of Economic Growth,* New York 1960.

"The No-Growth Society," *Daedalus,* American Academy of Arts and Sciences Proceedings, *102,* Fall 1973.

Resources as a Constraint on Growth

By WILLIAM D NORDHAUS*

For a considerable part of its history, the American economy has functioned as a cowboy economy. It has been a cowboy economy in the sense that there have been no important resource constraints on growth. This is not to say that land, minerals, and a clean environment have been freely available. Rather, agricultural land could be obtained at roughly constant costs; most essential minerals have been present at fairly high grade in considerable abundance; and the environment could be used as a sink without becoming fouled. In the last several decades, however, cropland has stayed almost constant. Some high grade mineral deposits have been exhausted, and the carrying capacity of our environment has been strained.

The scarcity of resources has led many to argue that the operating rules of our economy must change. Whereas in the cowboy economy we could afford to use our resources profligately, the new view of economic growth is that the closing of all our frontiers means that we are now operating in a spaceship economy. In a spaceship economy, great attention must be paid to the sources of life and to the dumps where our refuse is piled. Things which have traditionally been treated as free goods—air, water, quiet, natural beauty—must now be treated with the same care as other scarce goods.

It would seem difficult to question the observation that the world economy is progressing toward a closed system. Many have carried this observation further, describing a future imperiled by famine, depleted of essential materials, running out of energy, or choking in its own exhaust fumes. Behind these pessimistic visions is a deeper skepticism about the very fruits of economic growth.

Economists have for the most part ridiculed the new view of growth, arguing that it is merely Chicken Little Run Wild. I think that the new view of growth must be taken seriously and analyzed carefully.

What have we learned about the new view?

The first set of studies relates to *theoretical* investigations. (By theoretical I mean propositions based on largely untested assumptions about model structure —perhaps hypothetical would be a more accurate term.) In this category belong the celebrated writings sponsored by the Club of Rome as well as many offshoots of this work (Jay Forrester, Dennis H. Meadows et al.). These works have demonstrated that, under certain conditions involving technology, population, and resource availability, a sustained growth path for consumption is not possible.

The conclusions of these works have not generally been accepted by economists because of the dubious nature of many of their assumptions. In particular, the assumptions regarding population growth and technology are quite unsatisfactory. Several authors have shown that the conclusions of these models are not robust to minor modifications in structure. Thus R. Boyd showed that introducing a new factor called "technology" would drastically alter the model's path. My work (1973a) showed that any of three changes in model structure—ongoing technological progress, adequate factor substitution or population decline—would lead to opposite and more optimistic results.

It should be stressed, however, that all of the debate about Club of Rome models has been theoretical. It has been demonstrated that different world models paint drastically different pictures of future economic life. It is at this moment an open question as to which is the preferred model. Only careful empirical analysis can indicate which of the alternative models is closer to reality.

---

* Yale University.

What then of *empirical studies* of long-run constraints on growth? Although there have been no comprehensive studies, several particular problems have been investigated. I will report on recent findings for mineral resources and energy.

For centuries there has been virtually constant concern about the availability of mineral resources. Many recent studies have kindled this anxiety by showing that the ratio of proved reserves to current consumption (*R–C* ratio) for most minerals is very low. Fortunately, this is unduly pessimistic because the concept of reserves is entirely different from ultimate recoverable resources. Proved reserves are akin to the working capital or inventory of known resources. At the opposite extreme it is possible to calculate the total crustal abundance (*CA*) of different materials; of course this is unduly optimistic because it assumes that everything can be recovered. Somewhere between the two concepts lies the economically relevant measure—ultimate recoverable resources (*URR*). Although *URR* is a variable which depends on technology and price, we can as a first approximation use recent estimates prepared for the U.S. Geological Survey. These assume that *URR* is approximately 0.01 percent of total availability to a 1-kilometer depth (*Geological Survey*, p. 23). It should be emphasized that these estimates do not take into account the economic feasibility of mining lower grade ores as prices rise or techniques improve.

Table 1 shows estimates of the abundance of eleven important resources according to these three measures. (The first ten are the most important minerals by value of production.) It should be noted that *URR* is the most uncertain, especially because it involves estimates of what future prices and technologies will be.

The clear evidence is that the future will not be limited by sheer availability of important materials; rather, any drag on economic growth will arise from increases in costs. This raises the question of long-run movements in extraction costs of different materials.

TABLE 1—RESOURCE AVAILABILITY FOR IMPORTANT MINERALS BY THREE MEASURES

	Known Reserves/ Annual Consumption (*R/C*)	Ultimate Recoverable Resource/Annual Consumption (*URR/C*)	Crustal Abundance/ Annual Consumption (*CA/C*)
Coal	2,736	5,119	na
Copper	45	340	242,000,000
Iron	117	2,657	1,815,000,000
Phosphorus	481	1,601	870,000,000
Molybdenum	65	630	422,000,000
Lead	10	162	85,000,000
Zinc	21	618	409,000,000
Sulphur	30	6,897	na
Uranium	50	8,455	1,855,000,000
Aluminum	23	68,066	38,500,000,000
Gold	9	102	57,000,000

Source: Geological Survey, pp. 22–23, 613–14, 140. *Statistical Abstract*, p. 651.

It is useful to view the problem of rising materials costs in terms of a simple model of production. If interest rates are relatively constant, then we can express the costs of production in terms of the costs of two primary factors, labor and resources. (Capital is simply dated labor and resources.) One way to view the long-run problem is whether technological progress has been sufficiently rapid to offset the diminishing returns in resource extraction which arises from mining lower grade ores (or deeper or thinner veins). A simple index of this process is the movement of the "labor cost of resources," that is, the ratio of resource price to labor price.

Table 2 shows the ratio of the prices of the eleven most important minerals to the price of labor. This indicates that there has been a continuous decline in resource prices for the entire century. Unless all minerals suddenly hit a kink in the cost curve at the same time, it seems unreasonable to foresee a drastic runup of the cost of minerals relative to wages in the near future.

In summary, we expect that higher consumption levels in the future will lead to mining of lower and lower grade ores. Whether or not this leads to continuing decline of the resource/labor price ratio depends on whether technological progress continues to outstrip the movement to lower grade ores.

TABLE 2—RELATIVE PRICE OF IMPORTANT MINERALS TO LABOR, 1970=100

	1900	1920	1940	1950	1960	1970
Coal	459	451	189	208	111	100
Copper	785	226	121	99	82	100
Iron	620	287	144	112	120	100
Phosphorus	—	—	—	130	120	100
Molybdenum	—	—	—	142	108	100
Lead	788	388	204	228	114	100
Zinc	794	400	272	256	126	100
Sulphur	—	—	—	215	145	100
Aluminum	3,150	859	287	166	134	100
Gold	—	—	595	258	143	100
Crude Petroleum	1,034	726	198	213	135	100

Source: Values are the price per ton of the mineral divided by the hourly wage rate in manufacturing. Data are from *Historical Statistics, Long Term Economic Growth, Statistical Abstract.*

A specific class of resources where there has been considerable concern and quantitative work is energy resources. Energy is in fact an excellent case study of the central propositions about the limits to growth. Unlike most mineral resources, energy is essential for many processes of production. In addition, energy resources are finite in supply; energy consumption is dissipative, and currently leads to serious environmental problems.

Three important questions about energy resources are important for assessing future long-run growth. These concern resource availability, price and environmental effects.

The first question is whether there are adequate energy resources to run the world's economy for an indefinite period of time. Unfortunately, a complete answer depends on the answers to the next two questions, but for a crude answer we can simply calculate the quantities of energy resources. Table 3 shows the ratio of reserves to 1970 consumption under certain assumptions about the feasible technology.

Pretty clearly, the sheer adequacy of energy resources depends on whether certain future technologies will become available. Even with only the current technology (Table 3, line 2) there are resources for more than 8,000 years at the current rate

TABLE 3. RESOURCE-CONSUMPTION RATIO FOR ENERGY, 1970

1. Fossil fuels only	520
2. Fossil fuels plus current nuclear technology	8,400
3. Fossil fuels, current nuclear, and breeder technology	1,100,000
4. Fossil fuels, current nuclear, breeder, and fusion technology	53,000,000,000

Source: Hubbert, *Statistical Abstract* 1972, Nordhaus (1973b).

of consumption. With breeder reactors, and more dramatically with a fusion technology, there is virtually unlimited energy available.

The second question concerns the price at which energy resources will be available over the long run. The new view of growth sees a future in which energy resources become extremely expensive as mining turns to lower and lower grade energy resources. Thus it may well be that resources are available, but they are so expensive and so many resources are devoted to extraction that a dwindling amount of output is left for actual consumption.

I have recently considered the likelihood of this in another context (Nordhaus, 1973b). Starting with energy resources, with plausible demand paths, and with current estimates of the costs of extracting and processing energy resources, it is possible to calculate the results of a stimulated competitive market unfolding over time. In making this calculation, I assumed that the world's energy resources are efficiently allocated over time and space with an interest rate of 10 percent, and that the supply of energy resource was that given by technological assumption 3 in Table 3 (that is, breeder reactors would be technically and environmentally feasible by the year 2010).

The calculated path for energy prices was quite remarkable. The starting point in 1970 showed calculated energy prices very close to the actual. The efficient solution traced out a transition to the ultimate technology, with the final nuclear-based technology ultimately taking over about 150 years out. The most interesting part was the price path. Over the next 50 years, the calculated price index of final energy products rose 2.2 percent annually relative to the general price level. The growth rate of energy prices relative to the general price level over the next 100 years was 1.3 percent annually—after which time energy prices were constant relative to the general price level.

Whether movements of relative prices of this magnitude become a drag on consumption depends on the future pattern of technological change. If productivity increases at its historical rate—output per manhour growing at about 2.5 percent over the last few decades (see E. Denison)—then a rise in energy prices of the order of magnitude cited above would mean that energy prices would continue to fall relative to labor's price and to average incomes. In this case, there would be no drag on consumption standards stemming from the gradual exhaustion of low cost fuels. Under the assumptions of the model it thus appears that the long-run outlook for energy prices is favorable, although less favorable than over the last few decades. (The assumption of competitive behavior is clearly unrealistic. A cartel of oil producers would drive up the short-run price. But as Robert Solow has noted, monopolists are the conservationists' best friends: higher prices lead to lower consumption, a stretching out of finite resources, and possibly even lower prices in the future.)

The final and probably the most difficult question concerns the environmental effects of energy use. Up to now, there has been considerable attention given to the "local" environmental problems—especially sulphur emissions from stationary sources and the emission of oxides of nitrogen, carbon monoxide and hydrocarbons from automobiles. Adapting a technology to these new constraints has proved difficult and costly, but there seems to be wide agreement that—with sufficient time and money—emissions can be brought in conformity with any reasonable set of standards (see National Academy of Science Report).

A second set of environmental problems

concerns "global" standards, of which the most prominent is the earth's heat balance. The report on *Man's Impact on the Global Environment* reviews very thoroughly knowledge about the possible adverse global effects of energy consumption. The most important of these for energy were a possible "greenhouse effect" stemming from a large increase in carbon dioxide (CO_2) production from fossil fuels and the problem of global heat balance. Although there is great uncertainty, meteorological studies indicate that a change of two degrees Centigrade in average temperature is the order of magnitude which could trigger "albedo instabilities" leading to melting ice caps or "equator-to-pole ice cover" (Massachusetts Institute of Technology, p. 98). Studies report that such a temperature rise could result from a doubling of atmospheric CO_2 or from net waste heat of around 3 percent of total solar radiation (*MIT*, p. 88). These thresholds are well beyond current levels of human activity. Moreover, our limited supply of fossil fuels limits the production of CO_2 to acceptable levels.

I have performed a rough calculation of the atmospheric concentration of CO_2 along the efficient path described above. Assuming that 10 percent of the atmospheric CO_2 is absorbed annually (G. Skirrow), the concentration would be expected to rise from 340 ppm in 1970 to 487 ppm in 2030—a 43 percent increase. Although this is below the fateful doubling of CO_2 concentration, it may well be too close for comfort.

Waste heat could conceivably be a problem—but not for a while. Energy consumption is currently about 0.003 percent of incident solar energy—a five hundred-fold increase (160 years at a 4 percent annual growth rate) would be environmentally unacceptable.

REFERENCES

R. Boyd, "World Dynamics: A Note," *Science*, Aug. 11, 1972.

E. Denison, *Why Growth Rates Differ*, Washington 1967.

J. Forrester, *World Dynamics*, Cambridge, Mass. 1971.

H. E. Goeller, *Proc. Nat. Acad. Sci. USA*, Oct. 1972, *69*, 2991–92.

———, *A Mineral Resources Primer*, unpublished manuscript, 1973.

W. K. Hubbert, "Energy Resources," in Nat. Acad. Sci. USA, *Resources and Man*, San Francisco 1969, 157–242.

D. H. Meadows et al., *The Limits to Growth*, New York 1972.

W. D. Nordhaus, "World Dynamics: Measurement without Data," *Econ. J.*, Dec. 1973a.

———, "The Allocation of Energy Resources," *Brookings Papers Econ. Act.*, 1973a, *3*.

G. Skirrow, "The Dissolved Gases—Carbon Dioxide," in J. P. Riley and G. Skirrow, eds., *Chemical Oceanography*, London 1965.

Massachusetts Institute of Technology (*MIT*), *Man's Impact on the Global Environment*, Cambridge, Mass. 1970.

National Academy of Science, *Report by the Committee on Motor Vehicle Emissions*, Feb. 15, 1973, Washington.

United States Department of Commerce, *Historical Statistics of the United States*, and *Statistical Abstract, 1972*.

U. S. Geological Survey, *United States Mineral Resources*, Geological Survey Professional Paper 820, Washington 1973.

A Growth Model with Population Endogenous

By Richard F. Kosobud and William D. O'Neill*

In this study we consider the stability conditions and equilibrium properties of a long-run model of economic growth in which population and labor force are endogenous. Our way of formulating the stability conditions of this model in terms of deviations from a trend solution enables us to relate the model to a variety of historical data and to begin to explore the disequilibrium properties of the model. An additional purpose of this study is to show how sensitive the model is to specification changes. The prime example of this is our finding that the stability of the model with endogenous population and labor force requires the rates of growth of the birth rate and per capita income to be negatively related.

I. The Model with Population and Labor Force Exogenous

We first follow the well-known path cleared by Robert Solow (1956) and others by specifying a model in which population and labor force are exogenous. We describe a developed, capitalistic economy with the following production function and savings relationships:

$$(1) \qquad Y(t) = \lambda(t)[L(t)]^{\alpha}[K(t)]^{1-\alpha} \qquad (0 < \alpha < 1)$$

and

$$(2) \qquad \dot{K} = s(t)Y(t).$$

Thus we assume a constant-returns-to-scale Cobb-Douglas production function in which flows of homogeneous labor and capital inputs, L and K, determine net output of a homogeneous good, Y. λ, which may vary over time, reflects the effects of technical progress; s, the fraction of income (net output) not consumed, may also vary over time. We also assume that employers equate marginal products to wage and rental rates, so that price adjustments occurring behind the scenes of our model will assure that employers will absorb labor and capital supplies.

In asserting that production is subject to constant returns to scale, we assume away problems of exhaustible resources and nondestructible but nonaugmentable land raised by James Tobin and William D. Nordhaus. In defining output as a homogeneous product, we assume that the economy has but one sector unlike the study of Hirofumi Uzawa. By introducing technological progress the way we do, we assume it to be disembodied and neutral, i.e., not affecting the relative marginal products, cf. Solow (1960). In addition, we do not consider whether changes in the savings ratio and the functional or size distribution of income are related (cf. Nicholas Kaldor and James A. Mirrlees), although we do permit changes in s to affect equilibrium in the model. The generality of our model loses something from these assumptions; but we indicate below several ways in which our conclusions might be modified in view of work which does not make these assumptions, and these losses must be balanced against the gains obtained from adding to a transparent model the assumption that population and labor force are endogenous.[1]

---

* Professors of economics and systems engineering, respectively, at the University of Illinois at Chicago Circle. Very helpful comments are acknowledged from Tjalling C. Koopmans and Mark R. Killingsworth, but we absolve them from any errors remaining. Research aid from the National Institute of Child Health and Human Development is gratefully acknowledged.

It will be convenient to work with the capital output ratio, $x(\mathrm{t})=K(\mathrm{t})/Y(\mathrm{t})$. Taking logarithms of this definition and differentiating with respect to time yields:

$$\dot{x} = x(\dot{K}/K - \dot{Y}/Y).$$

Expressions for the rates of growth within the parentheses may be obtained from (1) which when substituted into the above yield a first order, linear, nonhomogeneous differential equation in x with a time varying coefficient.

$$(3) \quad \dot{x} = -(\dot{\lambda}/\lambda + \alpha\dot{L}/L)x + \alpha s(\mathrm{t}).$$

In general (3) has an equilibrium, x^0, that can be found by setting $\dot{x}=0$. Thus, for $\dot{\lambda}/\lambda+\alpha\dot{L}/L\neq 0$,

$$x^0 = \alpha s(\mathrm{t})/[(\dot{\lambda}/\lambda) + \alpha\dot{L}/L].$$

In equilibrium, since $\dot{x}^0=0$, several well-known and several not so well-known results are demonstrable.[2]

(a) In constant equilibrium, λ, L, and s are not independent inputs. It is easy to show that,

$$s(\mathrm{t}) = \frac{s(0)\,Y(0)}{\dot{Y}(0)}\,(\dot{L}/L + \dot{\lambda}/\alpha\lambda).$$

which indicates the relationship holding among these variables, only two of which are free to vary. We assume $\mathrm{t}=0$ is the time in which equilibrium is achieved and maintained.

(b) In balanced growth or constant equilibrium, the level of $Y(\mathrm{t})$ is independent of λ and L and depends only on $s(\mathrm{t})$. The rate of growth of income (and capital), however, depends on $\dot{L}/L$ and $\dot{\lambda}/\alpha\lambda$ but not on $s(\mathrm{t})$.

Results (a) and (b) obtain under the rather severe assumption that the capital-output ratio has a constant equilibrium.

Although the notion that x is constant is widely accepted, some evidence (cf. Lawrence R. Klein and Kosobud) suggests that it has in fact fallen over time. To make our analysis more general, we therefore assume that at any point in time the actual capital-output ratio, $x(\mathrm{t})$, consists of its long-term trend value, $X(\mathrm{t})$—which may be changing over time or, in a special case, is constant—and a deviation from the trend $\xi(\mathrm{t})$. Thus $x(\mathrm{t})=X(\mathrm{t})+\xi(\mathrm{t})$. We further suppose that such deviations in the capital-output ratio are caused by deviations, $u(\mathrm{t})$, in the actual savings ratio, $s(\mathrm{t})$, from its long-term trend value, $S(\mathrm{t})$: thus, $s(\mathrm{t})=S(\mathrm{t})+u(\mathrm{t})$, and $S(\mathrm{t})$ may vary over time (ξ and u could be stochastic terms).

What are the stability properties of this kind of economy? It turns out that the deviation of the economy's actual capital-output ratio from its long-term trend value will be smaller, the faster technical change and the labor force are growing. To see this, note that substituting into (3) enables us to write for the perturbation of the trend solution,

$$\dot{\xi} = (-\dot{\lambda}/\lambda - \alpha\dot{L}/L)\xi + \alpha u.$$

What are sufficient conditions such that $\xi(\mathrm{t})$ is bounded and $\xi(\mathrm{t})\to 0$ as $\mathrm{t}\to\infty$ when $u(\mathrm{t})\to 0$? Or equivalently, what are the conditions such that $x(\mathrm{t})\to X(\mathrm{t})$ as $\mathrm{t}\to\infty$ when $u(\mathrm{t})\to 0$?

A set of sufficient conditions is:

(a) $|X|\leq X_m$; $|u|\leq U_m$. The trend solution and savings disturbance are bounded for all time.

---

[1] Further work is presented in Joseph E. Stiglitz and Uzawa, which contains additional references.

[2] For purposes of exposition and space constraints we have excluded rigorous proofs of our results Some of the proofs are quite easy while others are more subtle. The interested reader will find our proofs as corollaries and extensions of the work of C. A. Desoer (1970) and Desoer and K. K. Wong (1968). To demonstrate (a) note that $\dot{x}^0=0$. Thus, $\dot{s}/s=[d/dt(\dot{\lambda}/\lambda+\alpha\dot{L}/L)]/(\dot{\lambda}/\lambda+\alpha\dot{L}/L)$. Solving this equation and noting that $\dot{Y}(0)/Y(0)=(1/\alpha)(\dot{\lambda}/\lambda)+\dot{L}/L$ gives the result.

(b) There is a positive constant c such that for all $t \geq 0$,

$$\dot{\lambda}/\lambda + \alpha\dot{L}/L \geq c.$$

Remarks:

(1) With these assumptions we can show that $\xi(t)$ is bounded for all time. In particular it will be the case, taking $\xi(0)=0$, that

$$|\xi(t)| \leq (\alpha U_m)/c.$$

(2) An economic interpretation of c is possible. If λ and L are growing at constant rates, say n_λ and n_L then $c=n_\lambda+\alpha n_L$ so that

$$|\xi(t)| \leq (\alpha U_m)/(n_\lambda + \alpha n_L).$$

Thus the faster λ and L are growing the smaller is the deviation away from the trend X for a given disturbance in $s(t)$.

It is appropriate at this point to ask about the speed of equilibrium convergence (R. Sato). Given an assumption of constant rates of growth of λ and L (and a constant saving rate s_0), it is simple to write the equilibrium solution of (3) as

$$x_1 = \alpha s_0/(n_\lambda + \alpha n_L).$$

Suppose further that at $t=0$, $s(t)$ changes so that for $t>0$, $s(t)=s_0+\Delta s$ where Δs is a constant. This change will establish a new equilibrium for (3) at

$$x_2 = \alpha(s_0 + \Delta s)/(n_\lambda + \alpha n_L).$$

We answer the question of how long it takes for (3) to go from x_1 to x_2. Clearly $x(t)$ traverses the distance $\Delta x = x_2 - x_1 = \alpha\Delta s/(n_\lambda+\alpha n_L)$. The time to convergence to the new equilibrium is the time $\hat{t}$ satisfying for any $0<\delta<1$ the expression $(x(t)-x_1)/\Delta x=\delta$ which is easily calculated from the solution of (3) for $x(t)$. This is:

$$\hat{t} = [ln(1/(1-\delta))]/(n_\lambda + \alpha n_L). \tag{4}$$

For example, if $\delta=0.9$; that is, $x(t)$ is within 10 percent of its new equilibrium, $n_\lambda=0.02$; $n_L=0.05$; and $\alpha=0.84$; then, $\hat{t}=37$ time periods. It is important to note what this time period depends or does not depend upon:

(a) $\hat{t}$ is independent of s; and,

(b) the larger n_λ, n_L, and α, the smaller $\hat{t}$. Consequently, the more rapid is technological progress and labor force growth and the larger is output response to labor input increase (labor's share under competition), the more rapidly will the system converge to a new equilibrium. Similarly, if changes in n_λ or n_L are the reasons for a new equilibrium, we may obtain expressions differing from (4) only by the additive inclusion of Δn_λ or Δn_L in the denominator. We conclude in this instance that the system will converge more rapidly the greater the magnitude of the change is in the respective growth rates of either λ or L as long as these changes are positive.

The fact that the time to convergence is independent of $s(t)$ might be termed one aspect of the disequilibrium behavior of the model. However, the time to convergence depends crucially upon the form of the differential equation as well as the values of the variables and parameters. It is a possibility that new equilibrium position(s) will be formed more rapidly over time than the model can move toward the old one(s).

II. The Model with Population and Labor Force Endogenous

Making population growth depend upon the economic sector of our model brings about important changes in the properties of the extended system and one of these changes, at the least, proved to be a surprise to us. This was the requirement (for stability) that birth rate growth and per capita income growth be inversely related. In an effort to take advantage of the growing body of empirical work on fertility, we have chosen to make the growth rate of births depend upon per capita income growth rather than, in the more conventional tradition, to relate labor force growth to the real wage rate or per capita income. At the same time we have chosen to make the specification as simple as possible so as to see more clearly the consequences for the model of this interdependence. Our key behavioral relationship is:

$$\dot{b}/b = a[(\dot{Y}/Y) - (\dot{P}/P)], \tag{5}$$

where b is births per capita, P is population, and the slope coefficient, a, represents the influence of per capita income growth upon the growth of the birth rate. Simple as equation (5) is, it results in a nonlinear differential equation in the capital-output ratio much more difficult to analyze than (3). It is also an implication of (5) that the system exhibits stability only if a is negative, but not too negative. Such a result would appear to have significance for empirical work (and we are not at all adverse to drawing such implications for empirical work), but before our argument is interpreted to mean that children are "inferior goods," further interpretation is in order.

Equation (5) does not, of course, tell the whole story about the number of children families have. In reviewing current work on fertility, we have been impressed with the attempts to distinguish more subtle aspects of income contained within the relation of crude income measures to number of children, and to disentangle the effects on the number of children of other variables which may be highly related to income. In the latter category would fall such important variables as fertility control methods which are likely to be correlated with income, and as their effect is to reduce births, they may confuse the sign of the income-birth relationship. In the former category, it is important to distinguish the wife's income or employment opportunities from the husband's, as increases in the wife's earnings change the opportunity costs of having children more significantly, perhaps, than increases in the husband's earnings. However, this latter point is not at all clearly established. Further, growth in income from an increase in real wages, since it may affect the "cost of time" to the parents for the time-consuming activity of child raising, may result in a reduction in the number of children desired even though the income effect of the increase acts the other way. To the extent this distinction is empirically important, it would be necessary to separate, for example, income received from wages and salaries and income received from property and to estimate the effect of each income type upon the number of children.[3] These considerations point toward elaboration of the specification of (5) and caution us not to give a too narrow an interpretation.

Having qualified our interpretation of (5), it remains to be restated that the restriction on the sign and magnitude of a (for stability) arises from the specification of the entire model and not from the results of empirical work. It is of considerable interest, therefore, to demonstrate why those bounds are required. As a further incentive to this demonstration, we note that in the review of current work by B. Janowitz and in a recent study by Y. P. Venieris, F. D. Sebold, and R. D. Harper, the relation of the number of children to income is found to be negative in a manner not unlike relation (5). Therefore, there is some recent quantitative evidence in support of further study of the extended model.

As we are working with births per capita, it is necessary to append the following definitional relations to our system:

---

[3] We are indebted to Mark R. Killingsworth for making this point. We have completed a review of empirical fertility relationships and concluded that a number of them have implausible implications for long-run growth models. We intend to report on this survey at a later time.

$$\dot{P} = (b - d)P,$$

and

$$L = \beta P$$

which first define population growth (d is the death rate per capita assumed constant, and migration is assumed away). Labor force, L, is some fraction, β, of population, a fraction which can be time varying but lies between 0 and 1. Noting that $\dot{L}/L = (\dot{\beta}/\beta) + (\dot{P}/P)$ and making use of (3), we obtain an expression for birth rate growth:

$$\dot{b}/b = (a\dot{\lambda}/\alpha\lambda) + (a\dot{\beta}/\beta) + [a(1 - \alpha)/\alpha](\dot{x}/x).$$

From this expression we can find solution values for the birth rate as follows:

$$b = b(0)[(\lambda/\lambda(0))^{1/\alpha}(\beta/\beta(0))(x/x(0))^{(1-\alpha)/\alpha}]^{a}.$$

Denoting by C the integration constants on the right hand side of the above expression, the equation for the birth rate is as follows:

$$b = C\lambda^{(a/\alpha)}\beta^{a}x^{a(1-\alpha)/\alpha}.$$

This result together with the definition $\dot{L}/L = (\dot{\beta}/\beta) + (\dot{P}/P) = (\dot{\beta}/\beta) + (b - d)$ may be substituted into (3) to obtain a differential equation in the capital-output ratio embodying the dynamic properties of our extended model. This equation is:

$$\text{(6)} \qquad \dot{x} = -((\dot{\lambda}/\lambda + \alpha\dot{\beta}/\beta) - \alpha d)x - \alpha(C\lambda^{a/\alpha})\beta^{a}x^{\rho} + \alpha s.$$

Here $\rho = a(1 - \alpha)/\alpha + 1$. Our first project is to find sufficient conditions such that the trend solution $X(\mathrm{t})$ of (6) is stable in the sense that $\xi(\mathrm{t})$ is bounded and $\xi(\mathrm{t}) \to 0$ as $\mathrm{t} \to \infty$ when $u(\mathrm{t}) \to 0$ as $\mathrm{t} \to \infty$, where $\xi(\mathrm{t})$ and $u(\mathrm{t})$ are defined as before.

A set of sufficient conditions are:

(a) $|X| \leq X_m$; $|u| \leq U_m$.

(b) There is a positive constant c such that for all $\mathrm{t} \geq 0$,

$$\dot{\lambda}/\lambda + \alpha\dot{\beta}/\beta \geq c > \alpha d.$$

(c) $\alpha/(\alpha - 1) < a < 0$.

Remarks:

(1) Once again we can interpret the constant c as $c = n_{\lambda} + \alpha n_{\beta}$ where n_{β} is the constant growth rate of labor participation. Thus we require the rate of technological progress and the weighted participation rate to exceed the death rate weighted by labor's output elasticity.

(2) The final sufficient condition, which we conjecture also to be a necessary one, sets bounds on the response of the birth rate to per capita income growth. For stability, this response must be negative as in the Janowitz study but not less than $\alpha/(\alpha - 1)$. Directing our attention to the interpretation of this latter term, we note that as α increases toward unity the response of income growth to a percent change in labor approaches unity, tending to equalize the growth rates of income and population, *ceteris paribus*, and permitting the slope coefficient a to assume smaller (absolutely larger) values without impairing the stability of the system.

This not intuitively obvious result together with questions of disequilibrium behavior of the extended model indicates the nature of some of the research that lies ahead. It will also be of great interest to extend the birth rate dependence upon income along the lines of our prior discussion and ascertain the sensitivity of the model's properties to these specification changes.

REFERENCES

C. A. Desoer, "Chapter VI" in *Notes for a Second Course in Linear Systems*, New York 1970.

——— and K. K. Wong, "Small-Signal Behavior of Non-Linear Lumped Networks," *Proceedings* of the I.E.E.E., Jan. 1968, *56*, 14–22.

B. Janowitz, "Cross-Section Studies as Predictors of Trends in Birth Rates: A Note on Ekanenis' Results," *Demography,* Aug. 1973, *10,* 479–481.

N. Kaldor and J. A. Mirrlees, "A New Model of Economic Growth," *Rev. Econ. Stud.,* June 1962, *29,* 174–192.

L. R. Klein and R. F. Kosobud, "Some Econometrics of Growth," *Quart. J. Econ.,* May 1961, *75,* 173–198.

R. Sato, "Fiscal Policy in the Neo-Classical Growth Model: An Analysis of Time Required for Equilibrating Adjustment," *Rev. Econ. Stud.,* Feb. 1963, *29,* 16–23.

R. M. Solow, "A Contribution to the Theory of Economic Growth," *Quart. J. Econ.,* Feb. 1956, *70,* 65–94.

———, "Investment and Technical Progress," in K. J. Arrow, et al., eds., *Mathematical Methods in the Social Sciences,* Stanford 1960.

J. E. Stiglitz and H. Uzawa, *Readings in the Modern Theory of Economy Growth,* Cambridge, Mass. 1969.

J. Tobin and W. Nordhaus, "Is Growth Obsolete?" in *Economic Growth,* New York 1972.

H. Uzawa, "On a Two-Sector Model of Economic Growth II," *Rev. Econ. Stud.,* June 1963, *30,* 105–118.

Y. P. Venieris, F. D. Sebold, R. D. Harper, "The Impact of Economic, Technological, and Demographic Factors on Aggregate Births," *Rev. Econ. Statist.,* Nov. 1973, *60,* 493–497.

POPULATION AND ECONOMIC POLICY

Resource and Amenity Implications of Population Changes

By RONALD G. RIDKER*

Two observations are in order before I discuss the topic suggested by the title. First, interest in it forces us to be future oriented, and oriented towards the longer-term future at that. This interest, while arising from a variety of sources, generally involves predictions of growing scarcity of natural resources and environmental carrying capacity, worries about increasing dependence on foreign supplies and concerns about loss of local amenities due to the rapid influx of population into limited regions. It is not sufficient to study past trends when the overriding concern is that the future will not be like the past. Otherwise, we are subject to the criticism of those like E. F. Schumacher who recently suggested that economists spend most of their time optimizing the arrangement of deck chairs on the Titanic. Second, we cannot discuss the effects of population growth independent of other factors that can cause resource and environmental pressures. While it makes sense to ask about the relative importance of population in causing these pressures, we cannot completely ignore tastes, technology, institutions, policies, and international relations. These and other determinants strongly condition the resource and amenity pressures generated by population growth; they often change and affect the situation more than do demographic changes within specific time periods; and they often interact with each other and with demographic variables. In particular, we cannot set aside changes in per capita output since so many of the problems we are concerned with here are related to the movement of materials through the total economy.

These two considerations have led me to focus on three questions: Are there limits to economic growth imposed by resource exhaustion and environmental pressures? If so, how close are we to these limits? How important is population as opposed to other factors in moving us towards these limits? In discussing these questions I will consider primarily the United States.

The first question has significance only if the word "growth" is taken to mean material economic growth or growth in population and material economic output, but not growth in *GNP* or national income. Since the latter are artifacts in large part measuring the value placed on the goods and services produced by an economy, they could conceivably continue growing even without any growth in the amount of materials embodied in these goods and services.

Defined in this way, growth is quite definitely subject to limits. First, there are ultimate or natural limits imposed by the finite character of the earth. Ruling out the possibility of importing materials and energy from outside the earth on an ever-increasing scale, the second law of thermodynamics, the entropy law, makes the existence of this limit quite certain (see Nicholas Georgescu-Roegen). Scientific and technological advance can postpone and moderate the way this limit is reached but cannot repeal fundamental laws of nature. Ultimately, therefore, we must choose how we want population and material economic growth to cease, through the operation of natural forces or in a manner and at a time chosen by man.

---

* Program director, Resources for the Future, Inc.

Second, limits are imposed by "rate and magnitude" problems, by the speed with which, under the best of circumstances, the problems thrown up by growth can be resolved. These problems are related to the time lag inherent in research, investment and institutional change, lags that cannot be significantly reduced except perhaps at the cost of growth itself. A dramatic example of this kind of limit is suggested by the observation that, even if we built one large nuclear power station per day between now and the end of the century, fossil fuels would still have to meet more than half the world's primary energy requirements if these needs continue to grow at 5 percent a year (A. Lovins).

Third, a host of more proximate or man-made barriers to growth and development—institutional, political and international arrangements—restrict the ability of individual countries or groups within countries to take full advantage of the resource and environmental base available. Land ownership patterns that result in good agricultural land being withheld from intensive cultivation in densely populated, poor countries, cartel controls on oil sales and prohibitions on international migration are cases in point.

While these last two types of limits blend into each other somewhat, the distinction is useful. In principle, at least, the third type of limit can be eliminated, whereas with regard to the second there is probably some maximum degree of efficiency in problem solving and investment that cannot be surpassed at any one point in time without overloading the system.

That these limits exist is, it seems to me, clear but how close are we to them? It makes an enormous difference for policy today whether these ultimate limits are 50, 500 or 5,000 years away.

Of course, one study, *Limits to Growth*, argues that the ultimate limits are sufficiently close that we must immediately take action to stop world-wide population and economic growth (D. H. Meadows et al.). But as has been pointed out many times the model on which this study is based is far too aggregative, is too devoid of adjustment mechanisms, and incorporates far too many pessimistic assumptions —about technological progress, future reserves of nonrenewable resources, the ability to control, treat and absorb pollution and the extent of future population growth—to be taken seriously. This study and others—e.g., the work of Jay Forrester and essayists like Kenneth Boulding and E. J. Mishan—have been very important in forcing us to face the fact that there are limits to growth; but their estimates of the timing of such ultimate limits could easily be off by several centuries.

Far more persuasive are a series of studies, recently reviewed by a United Nations Symposium on Population, Resources and the Environment, which took a look at the physical magnitudes of minerals, energy, water and land and found them to be sufficient on a global scale to sustain current growth rates for a period well beyond any sensible planning horizon (see D. B. Brooks and P. W. Andrews, Lovins, and R. Ravelle). However, serious environmental problems could arise. For example, heat balances that control weather patterns could be disturbed, the accumulation of contaminants could disrupt the ecological services performed by the oceans and the spread of monocultures could result in crop failures on an ever-increasing scale (see Paul Ehrlich and J.

Holdren). But such assertions of threats to the ecosystem lack specificity: no one knows the magnitude of the risks we are running, how serious such breakdowns would be, or the likelihood of being able to avoid them through technological and other changes short of reducing growth itself. While we obviously need more information on such dangers, a strong case for the argument that we should immediately cease growth because these ultimate limits to growth are too close for comfort is difficult to make.

The situation with respect to the second and third sets of limits is quite different. Taken together they can explain the abundance of shortages that have befallen us in the last couple of years. Most of these shortages appear to have arisen from the world-wide economic boom conditions of the past few years, which have raised the demand for meats, timber, energy, fertilizer and other materials more rapidly than capacity can catch up, plus a series of special disruptions and managerial mistakes on the supply side, all of which have come at once and none of which is necessarily permanent in the sense of the ultimate limits.

While these limits, especially the second, have not been studied in any detailed way, the results of a recent study by Resources for the Future (*RFF*) for the President's Commission on Population Growth and the American Future shed at least some light on them (see Ronald Ridker). Using alternative assumptions about population and economic growth rates, this study asked whether the United States could meet its requirements for some twenty resources during the next quarter to half century by internal development or imports, and (under different assumptions about policy) estimated the amounts of fourteen pollutants forthcoming. Since it assumed no restrictions on international trade, no delays in the building of production facilities (resulting, for example, from conflicts over siting of nuclear power plants), no problems in implementing pollution controls, no shortages of skills, and so on, it comes closer to what is needed to assess the second rather than the third of these limits.

The results are not disheartening. At least for overall economic growth rates up to 4 percent and for population growth rates up to 1.5 percent, the highest rates considered, resource availability and environmental problems do not appear to pose any serious impediments to growth. Some costs will have to be paid and a number of social and institutional adjustments made, but no catastrophes are likely solely as a result of population and economic growth during the next half century. The resources, the technical knowhow and the possibilities for institutional accommodation are all present in sufficient degree to cope with whatever problems the particular resources and pollutants studied are likely to give rise to during this period.

But for present purposes this study is subject to a number of important qualifications. The study did not examine all pollutants and ecological threats and assumed that time lags for investment and the utilization of newer technologies will be the same as they have been in the past; it considered only a limited range of population and economic growth rates and a limited time horizon; it gave no weight to the strong possibility that, as population and material standards of living increase, social and technological interdependence and complexity will also grow, making society increasingly vulnerable to political and other breakdowns.

These qualifications suggest that we should be cautious about concluding that the second and third sets of limits, like the first, are too far away to worry about. No matter where a country is with respect to the ultimate limits, a higher growth rate means more rapid depletion of domestic and international resources, greater pressures on the environment, more dependence on continued rapid technological developments to solve these problems, less time to overcome social, institutional and political barriers to progress, fewer social options and a continued postponement of the resolution of the problems caused by past growth. In contrast a lower growth rate buys time, resources and additional options: time to find solutions, resources to implement solutions and additional freedom of choice in deciding among various patterns of living.

Moreover, if we combine these qualifications with the likelihood that, as income per capita grows, the desire for visual, climatic and related amenities increases faster than the demands for additional material goods, we could well find that we will simply not opt for a number of the technological and institutional changes that will become necessary to permit growth to continue at past rates. We may, for example, decide not to develop the fast breeder reactor, given all its environmental difficulties and related siting problems. Considerations such as these are more likely to limit growth than are physical limits, at least within the foreseeable future. It is unfortunate that *Limits to Growth* and similar writings have focused on ultimate limits rather than on limits of these other two types where more serious and relevant issues lie.

I now turn to the final question posed at the outset: How important is population growth relative to other determinants of resource and environmental pressures? Much of the discussion of this question has started with the notion that most resource and environmental problems arise primarily in conjunction with the flows and use of minerals, energy and other resources through the economy. Other things constant, this flow is closely related to the overall size of the economy. Since total output can be broken down into population and per capita output, a particular resource or environmental problem can be defined as equal to some function of the product of these two components of total output. A number of writers have used this simple notion, plus some data on indexes of these and related variables, to argue about the relative importance of growth in population versus growth in affluence and changes in technology (see, for example, Barry Commoner, Ehrlich and Holdren in Ridker).

But in any long-run analysis, other things do not remain constant. If technological advances make it possible to obtain the same output with less resources, or if tastes shift away from material goods towards services, the impact of a change in population or per capita *GNP* on resource requirements and the environment will be less. On the other hand, if land of poorer quality must be called into production to satisfy growing food requirements or if coal is again substituted for oil on a significant scale, the environmental problems associated with an increase in demand, whether from growth in population or income, will be greater.

The *RFF* study attempted to answer this question using several different models. By specifying time paths for the intervening variables and then running these models into the future under alternative assumptions about population growth, per capita *GNP* growth and changes in policy, and comparing the results obtained in the year 2000, we were able to reach the following conclusions. First, so far as resources and pollutants related to material flows through the economy are concerned, a 1 percent change in per capita *GNP* is more important than a similar change in population. For example, resource-use elasticities for population and per capita *GNP* are 0.56 and 0.85 for iron, 0.22 and 0.68 for tin, and 0.56 and 1.76 for sulfur, while pollution generation elasticities for comparable changes in these two factors are 0.29 and 0.59 for hydrocarbons, 0.42 and 0.87 for sulfur oxides, and 0.29 and 0.31 for nitrogen oxides (Ridker, pp. 45 and 47). In large part this is because, within the economic model utilized, a change in population induces an offsetting change in per capita *GNP*, whereas an exogenously introduced change in the latter does not react back on population growth. When the effect of the induced change in per capita *GNP* is factored out, the differences between the elasticities are substantially reduced.

But second, policies directly attacking resource use and pollution emissions tend to dominate those involving reductions in population and/or economic growth. The clearest indication of this is to be found for pollutants where we considered the effect of a technologically and economically feasible set of restrictions in emissions per unit of output on the total amount of pollution emitted in the year 2000. For example, while a shift from the high to the low growth rates in *both* population and per capita *GNP* considered in this exercise would reduce emissions of hydrocarbons in the year 2000 by 17 percent, the active abatement policy considered would reduce emissions by more than 80 percent, which is about a third less than we estimate them to have been in 1970. For all pollutants studied, such a policy would cost less than two percent of *GNP* in the year 2000, compared with about one percent in 1970, provided—an important qualification—that the total cost of the policy could be spread fairly smoothly over this thirty-year period. While the total cost is large, paying it would be equivalent to giving up less than one-tenth of 1 percent of growth in *GNP* over this period. Similar arguments can be made for resources given the extent of the substitutions available between them.

Third, for land- and water-based activities, the relative importance of population and economic growth is reversed, a decline in population growth generally reducing consumption and pollution emissions by more than a comparable reduction in per capita *GNP*. This appears to be the case for water use, land use, food production and even the use of a number of types of outdoor recreation facilities. But in these areas as well one can find a number of more direct ways to reduce consumption, pollution and congestion which are likely to be considered superior to across-the-board reductions in population and economic growth.

In conclusion—and by way of drawing an important policy implication from these considerations—this last point can usefully be emphasized. At least in the United States during the next quarter to half century, direct attacks on resource and environmental problems—for example, the use of effluent charges, zoning regulations, subsidies for research and the like—should be far more effective than attempts to accomplish the same thing by reducing population and material economic growth. In the process of applying such direct measures economic growth is likely to slow down somewhat, but by far less than it would take to accomplish a comparable reduction in a resource or environmental problem by an across-the-board reduction in growth. This, it seems to me, is the way that reductions in growth should occur if they occur at all. It is time that we dethrone growth not only as a goal but also as a source of evil and concentrate instead on the solution of specific problems. Of course, if such direct attacks prove insufficient, it would then be appropriate to consider more general limits to population and economic growth. A number of countries, particularly poor ones with fewer degrees of freedom, are likely to arrive at this point relatively soon, if they are not there already. But the United States still has time and other options; the problem for this country is to find ways to utilize its additional degrees of freedom wisely.

REFERENCES

D. B. Brooks and P. W. Andrews, "World Population and Mineral Resources: Counterintensive or Not?" U.N. Symposium on Population, Resources and Environment, Stockholm 1973.

B. Commoner, "The Environmental Cost of Economic Growth," in R. Ridker ed., *Population, Resources and the Environment,* Vol. III of Research Reports of The Commission on Population Growth and the American Future, Washington 1972.

P. R. Ehrlich and J. P. Holdren, "Human Population and the Global Environment," U.N. Symposium on Population, Resources and Environment, Stockholm 1973.

——— and ———, "Impact of Population Growth," in R. Ridker ed., *Population, Resources and the Environment,* Vol. III of Research Reports of The Commission on Population Growth and the American Future, Washington 1972.

N. Georgescu-Roegen, *The Entropy Law and the Economic Process,* Cambridge, Mass. 1971.

A. Lovins, "Energy Resources," U.N. Symposium on Population, Resources and Environment, Stockholm 1973.

D. H. Meadows et al., *The Limits to Growth,* New York 1972.

R. Revelle, "Will the Earth's Land and Water Resources Be Sufficient for Future Populations?" U.N. Symposium on Population, Resources and Environment, Stockholm 1973.

R. Ridker, ed., *Population, Resources and the Environment,* Vol. III of Research Reports of the Commission on Population Growth and the American Future, Washington 1972.

The Role of Population in Models of Economic Growth

By Allen C. Kelley*

More than a decade ago, Simon Kuznets accurately summarized the state of knowledge on the quantitative relationships that have been identified between demographic change and economic growth:

> . . . we have not tested, or even approximated, empirical coefficients with which to weight the various positive and negative aspects of population growth. While we may be able to distinguish the advantages and disadvantages, we rarely know the character of the function that relates them to different magnitudes of population growth. [p. 339]

Despite marked advances since then in the development of economic and empirical analysis on population, his appraisal lamentably holds with virtually the same force today.

In this paper we argue that most of the existing macroeconomic models which incorporate relationships between demographic and economic change are deficient in several dimensions. As a result, even the common finding that faster population growth significantly reduces rates of per capita output growth is typically based on macroeconomic frameworks whose empirical and theoretical foundations are shaky. We then suggest several ways in which economic-demographic models could be enriched. Finally, we present some recent results illustrating the potential usefulness of expanded models of economic-demographic interactions.

I. Previous Modeling

The most interesting macroeconomic-demographic model of low income countries is the pioneering framework provided by A. J. Coale and E. M. Hoover. In this model, rates of economic growth are in large part determined by the growth of the labor force and the amount of capital available per laborer. Population growth itself carries with it no positive impacts on capital formation rates. The basis of the adverse impact of population is straightforward. Faster growing populations imply more dependents, who consume but are assumed not to contribute to or stimulate an expansion in output. "Adult-equivalent" consumption increases and is financed totally out of saving. (Government outlays on young populations are also assumed to be relatively unproductive.) It is shown that projections of high population growth yield significantly lower levels of per capita (and adult consumer) output for India over the projection period.

The Stephen Enke *TEMPO* models, an excellent study by Paul Demeny, and Robin Barlow's analysis of the impact of malaria eradication in Ceylon are all in the same vein. These models possess several notable strengths. Each is rigorously formulated and the population-economic interactions are explicitly revealed. Each is simple and unambiguous. Each is efficient: relatively small demands on scarce data yield significant results.

In two basic areas, however, these models are deficient. First, their simplicity entails sacrificing important hypotheses

---

* Professor of economics and Associate Director, Center for Demographic Studies, Duke University. I am grateful to John Spitzer, who provided valuable research assistance in the preparation of this paper.

relating to economic-demographic interactions; and the economic-demographic hypotheses included are, in several instances, untested empirically and questionable theoretically. Second, the usefulness of these models for public policymaking is limited. The main policy implication is that slower population growth rates are desirable. But how is this best accomplished? What rate is appropriate? What is the impact of population growth on the structure of demand and production? On the rates of urbanization? On the rate and location of unemployment? Answers to such specific policy-related questions require more detailed frameworks than those hitherto developed.

II. Directions for Model Development

Given the research on population and economic development in the last decade, it is clear that the earlier pioneering models of economic-demographic growth may now benefit from revision and updating. Consider a few of the revisions which must rate high on a priority list for model development.

A. *Endogenous Population Growth*

Macroeconomic model building should specify population change as endogenous. A visible lesson from the historical record of the present advanced countries derives not from the impact of demographic change on rates of economic prosperity, but the responsiveness of population to economic change. Moreover, for a wide range of countries, considerable progress has been made in isolating relevant determining factors of desired family size. These factors have been deduced, in some instances, from well-articulated microeconomic specifications of household behavior. Recent research has also helped reconcile apparent divergences in population response to short- and long-run economic conditions, as well as the conflicting findings forthcoming from cross-section versus time-series studies.

Models in which population is endogenous are particularly relevant for two specific policy-oriented reasons. First, if population growth responds to economic conditions, then typical theoretical and empirical estimates of population's impact will be biased. For example, if population growth exerts a negative impact on the level and the pace of economic progress, and if family size varies inversely with economic conditions, then population growth will *decline* in response to its own hypothesized negative impact on the economy. Its *net* adverse impact on growth will therefore be quantitatively less than in the case where population growth is assumed to be exogenously given. In a recent study, A. C. Kelley, J. G. Williamson and R. J. Cheetham (1972a) show that in a general equilibrium dualistic model the feedback of economic progress on the pace of demographic change is critical to appraising both the short- and the long-run *quantitative* impact of population on the economy.

Second, if population growth is endogenous and sensitive to the variables used to assess its impact—pollution, economic growth, social stability, employment—then policy recommendations should be formulated in a framework analyzing population's *net* impact. The relevant research issues include the speed of adjustment of population growth to policy objectives—and the extent to which the adjustment level is in some sense "appropriate."

B. *The Impact of Children on Family Saving*

Possibly the most widely investigated adverse impact of population on economic growth is based on the hypothesized negative effect of population on the availability of savings for capital formation. Many strongly held convictions prevail;

little hard evidence is available. The most commonly postulated association between family size and saving centers on the dependency rate, or the adult-equivalency framework. Here the impact of children is negative. However, the validity of this approach is dubious.

Consider the underlying basis of the adult-equivalency argument. Children are characterized as exerting a notable negative influence on household saving since children consume and fail to produce. This characterization is not entirely accurate on theoretical grounds. The impact of children on saving should be viewed not in isolation but within the family unit, where their total influence is largely exerted. Under not implausible assumptions, the direction of the impact of children (or increased family size) on *family* saving rates is indeterminate. Consider just a few of the possibilities: children may substitute for other forms of consumption, they may contribute directly to household income, they may stimulate parents to work more (or less) and/or harder, they may encourage accumulation of certain forms of assets (e.g., education), and they may stimulate the amassing of estates. These impacts are likely to vary with the level of economic development, and they may be nonlinear (and even nonmonotonic) in family size for a given income level.

Hard evidence on the relationship between family size and household saving rates is almost nonexistent. While several empirical studies examine the impact of family size on the *distribution* of consumer expenditures, evidence on its impact on *aggregate* consumption, income and saving is scarce.

C. *The Simultaneity of Household Decision Making*

One of the more serious methodological restrictions of most macroeconomic-demographic models is their treatment of household decision making. They rarely, if ever, specify any microeconomic connections between several critical decisions: family size, saving, expenditure allocations, and workforce participation. A productive reformulation of these relationships would treat these several decisions as *interdependent* and estimate the parameters of the behavioral equations describing these decisions appropriately (e.g., by two or three-stage regression procedures). Attempts by scholars to move in this direction, clearly a compelling research strategy on both theoretical and empirical grounds, are almost nonexistent. The "New Home Economics" studies (see, for example, T. W. Schultz, ed.), which highlight interdependencies between family size and workforce behavior, represent one exception.

We have recently experimented (see Kelley 1973) with this type of household behavioral framework, postulating a household decision-making model where saving, workforce participation, and family size decisions are all interdependent. The results—based on a sample of 1,953 American households in the year 1889 and estimated by 2–3 stage procedures—were surprising. The magnitude and the direction of the relationship between changes in household savings and family size depends critically on the *origin* of the family size and saving changes. Indeed, the impact of increased family size on savings is shown to be negative, positive, or zero, depending on three alternative factors inducing family size changes. The policy implications of this finding are significant. Family planning programs can have quite different effects on economic development depending on what specific factor induces the reduction in family size.

D. *The Level and Composition of Government Expenditures*

Any policy-oriented model of economic-demographic change must possess a meaningful specification of the government sector. Most models either suppress the role of government or incorporate hypotheses of government behavior which are restrictive and unsubstantiated. One common approach is to assume a constant per capita demand for government expenditures. Alternatively, younger populations shift demand toward "unproductive" expenditures which are financed either at the expense of "productive" expenditures, or by an expansion in the size of the government sector.

Hard evidence is scarce on the determinants of the level and composition of government expenditures where demographic factors are explicitly taken into account. Most studies employ the convenient homogeneity assumption—government spending is determined by the level of *per capita* income. The population's age structure is almost never taken into account. Yet to assume that an increase in the dependency ratio necessarily results in a proportionate increase in unproductive government activity ignores the possibilities for substitution against other unproductive forms of government spending. Moreover, an increase in dependents plausibly influences other important factors—government revenues, prices of government services, and so forth.

In a recent exploratory study, we have utilized evidence from a cross section of countries to test hypotheses explaining the level and composition of government expenditures. When demographic factors are taken into account, previous theories of government production must be modified. This applies particularly to Wagner's Law, a widely tested model of government behavior.

E. *Model Sectorization: The Role of Demand*

The rate of economic development, commonly measured by the growth of per capita output, is also associated with the rate of structural change, e.g., industrialization and urbanization. Clearly, models to be used for policy purposes should go beyond aggregate relationships and incorporate relevant sectoral divisions. Consider two examples which focus on the urban-rural dichotomy.

What is the impact of population growth on the *composition* of demand as between urban and rural goods? Typically, the shift of activity from rural to urban production is explained by Engel's Law: food expenditures occupy a declining proportion of household budgets as income rises. Hollis Chenery challenges this position and argues that supply factors—changes in comparative advantage in production which derive from alterations in technology and factor endowments—are equally important. But the Engel curve is a restrictive specification even for analyses focusing on demand factors. And such analyses typically omit demographic factors, though systematic changes in the rate of population growth, in ages, in average family size, and in urbanization (internal migration) are all part of economic development. In a 1969 study (see Kelley 1969), we have shown that the shift in demand away from agricultural products proceeds more rapidly (a) the faster the rate of internal migration into urban areas, and (b) the more the benefits of economic growth are distributed to the urban population. If, as is likely, a rapid movement toward urbanization will also diminish population growth, the demand forces moving the economy toward modernization will be strengthened. This study supports the inclusion of demand-compositional, demographic influences in policy-oriented models.

A final example will suffice. Empirical studies and theoretical considerations support specifications of lower family size in urban rather than rural areas. Models possessing a rural-urban division can readily capture this hypothesis. In one such model, we have treated aggregate population growth as responsive to rates of structural change. The shift of population from rural to urban areas results in a reduction in the overall rate of population growth. Simulations reveal that even this highly simplified demographic specification alters the appraisal of population's impact on economic development (see Kelley, Williamson and Cheetham 1972a).

F. *Selected Theoretical Developments*

There is no difficulty in identifying negative effects of population. The diminishing returns to fixed factors and the dampening effect on the aggregate capital-labor ratio are two well-known connections. But when one moves from the simple neoclassical or Malthusian paradigms to more detailed theoretical structures, population's impact on the economy becomes more complex. Consider a few of the relatively unexplored possibilities.

First, what is the impact of population growth on the rate of technical progress? Typically the latter is assumed to be unresponsive to economic or demographic factors. While formal specifications of endogenous technical change are difficult to analyze, empirical explorations, together with qualifications of macroeconomic models, should at the very least take possible population-technical change interactions into account.

Second, what is the impact of population where savings are determined not only by the level of income, but also by its source? Growth-theoretic analyses commonly find that diminished population growth increases labor's income share. If the propensity to save out of labor income is less than that out of capital income, then the overall savings rate could fall.[1]

Finally, what is the quantitative importance in the intermediate run of population growth where technical change is in part embodied in new capital goods? Here the age distribution of the capital stock becomes important in determining its total productivity. The pace with which modern techniques can be incorporated into production varies directly with the rate of population growth. Given the critical role assigned to technical progress in explaining the rate of economic growth by economic historians, and by those applying the sources-of-growth approach, the role of population through its impact in a model of embodied technical change could well merit close attention.

III. Preliminary Results of Expanded Economic-Demographic Models

To illustrate how models such as the ones I have been discussing can enrich policy-oriented macroeconomic-demographic models of development, I shall describe in summary form the results of one recent experiment

As argued above, microeconomic specifications of the impact of family size on household saving should allow for the possibility that children are in part financed by reduced consumption of other household members, and that children may alter the workforce behavior of parents. The parameters of one such model and an alternative, the commonly employed adult equivalency framework, have been estimated with a large data file on American households (see Kelley, 1973), and then incorporated into the Enke *TEMPO* framework, a widely used macroeconomic model. The *TEMPO* model was then simulated under two different assumptions: no population growth reduction, or a reduction to half the initial rate. While lower population growth yielded higher per capita income levels at the end of the twenty-five year period under both the Kelley and the adult equivalency hypotheses, the benefits of lower population growth are 30 percent lower in the Kelley version than under the adult equivalency hypothesis.

---

[1] A mathematical analysis of this possibility is provided by Kelley, Williamson and Cheetham (1972a, ch. 3). A quantitative exploration of population's impact in a low-income economy using this theoretical approach is provided in Kelley, Williamson and Cheetham (1972b). While increased population growth *initially* diminished the rate of per capita output growth, this impact was rapidly dissipated over time since the initial capital-shallowing effect was largely offset by the positive stimulus to investment demand.

That such a simple change should produce such different results strongly supports Kuznets' earlier view that the quantitative impact of population change on economic development is far from being established. The time has come to enrich microeconomic specifications and statistical analyses by expanding the earlier pioneering models of economic and demographic change and updating their assessment of population's impact on the pace of economic growth and structural change.

REFERENCES

R. Barlow, "The Economic Effects of Malaria Eradication," *Amer. Econ. Rev.*, May 1967, *57*, 130–157.

H. Chenery, "Patterns of Industrial Growth," *Amer. Econ. Rev.*, Sept. 1960, *50*, 624–654.

A. J. Coale and E. M. Hoover, *Population Growth and Economic Development in Low-Income Countries*, New York 1958.

P. Demeny, "Demographic Aspects of Saving, Investment, Employment and Productivity," in *World Population Conference 1965*, New York 1967.

S. Enke, *Description of the Economic-Demographic Model*, Santa Barbara 1971.

A. C. Kelley, "Demand Patterns, Demographic Change and Economic Growth," *Quart. J. Econ.*, Feb. 1969, *83*, 110–126.

———, "Savings, Demographic Change and Economic Development," presented at the Population Association Meetings, New Orleans, April 1973.

———, J. G. Williamson and R. J. Cheetham, *Dualistic Economic Development: Theory and History*, Chicago, 1972a.

———, ———, and ———, "Biased Technological Progress and Labor Force Growth in a Dualistic Economy," *Quart. J. Econ.*, August 1972b, *84*, 426–447.

S. S. Kuznets, "Population Change and the Aggregate Output," in *Demographic and Economic Change in Developed Countries—A Conference of the Universities-National Bureau Committee for Economic Research*, Princeton 1960, 324–40.

T. W. Schultz, ed., "New Economic Approaches to Fertility: Proceedings of a Conference," *J. Polit. Econ.*, March/April 1973, *81*, S1–S299.

The Impact of Population Growth on Employment

By ALAN SWEEZY AND AARON OWENS*

This paper focuses on three widely held concerns about the impact of changes in population growth on employment. A society with zero (or slow) population growth would, it is feared, (a) have trouble maintaining full employment, (b) be hobbled by a labor shortage, and (c) suffer from a decline of creativity.

In its simplest form, the worry about full employment is based on the notion that to prosper business must have a growing number of consumers. This is clearly a one-sided view. As Joseph Spengler pointed out, consumers are also producers (p. 306). Business gains nothing from population growth as far as the balance between consumers and producers goes.

A more sophisticated argument concerns the impact of population growth on investment and saving. According to neoclassical theory, capital can be profitably invested in three different ways: (1) in providing additional workers with equipment of the same type already being used—it will not pay, however, to build such equipment unless there *are* additional workers to operate it; (2) in increasing the amount of capital per worker within the range of already known possibilities; (3) in introducing new technology. If accumulation continues at a given rate (i.e., the propensity to save remains unchanged), a decline in opportunities of the first type (what Alvin Hansen called "capital widening") would mean more of the accumulating capital would have to be used in the second and third ways. Unless there is a compensating increase in the rate of technological change, this would mean a more rapid decline in the rate of profit than would have occurred if part of the capital had been drained off in outfitting workers with equipment of already existing types. In the case of the less-developed countries it is generally accepted doctrine that rapid growth of population and the labor force will "waste" capital which could otherwise be used to raise the amount of capital per worker. The only difference between the less- and the more-developed countries is that we usually assume there are plenty of opportunities for raising the level of capital per worker in the former, whereas we think it at least possible in the developed countries that increasing the amount of capital per worker might cause a fall in the rate of profit to the point where capitalists no longer would be willing to maintain investment at a rate necessary for full employment, preferring, beyond that point, to hold money rather than to acquire real capital goods.

While the separate elements in this analysis are unexceptionable, difficulties arise in combining the long-run classical theory of capital, which is couched in real terms, with short-run Keynesian theory of money flows and money holdings. It seems paradoxical, for instance, that additional labor becoming available through population growth would serve to hold profits up while the existence of unemployed labor would not have the same effect. Without resolving these difficulties, economists have concentrated their attention on a different aspect of the relation between population growth and employment: consumer spending and saving. Most of them seem unaware, in fact, that labor supply could have anything to do with the problem.

---

* California Institute of Technology.

Gardner Ackley, for example, criticizes Hansen for not seeing that even without population growth expansion of income could provide a *market* for the product of a growing stock of capital (p. 511). But that misses the point. The problem, in the absence of population growth, is not the impossibility of providing a market but the lack of labor to operate the additional capital goods. John Cornwall and Allen Kelley simply ignore the relation between labor supply and capital stock, confining their attention to saving and spending on consumer durable goods. Even Hansen, though he emphasized the importance of capital widening as an outlet for saving, failed to make it clear that capital widening is possible only if the supply of labor is increasing. He thus left himself open to Ackley's criticism.

As far as consumer demand goes both Hansen and Cornwall think the effect of population growth is likely to be favorable: favorable, that is, in an economy which may have trouble maintaining an adequate flow of demand. In capital-short economies it would, of course, be just the opposite. Cornwall refines and elaborates Hansen's argument but his central points are essentially the same. He assumes that the propensity to spend on nondurables and services is unaffected by population growth but that "family related" expenditures (housing and other consumer durables) would be larger with a growing population. He also thinks the pressure to borrow is greater for couples making first purchases of consumer durables—especially housing—than for other types of consumers. Kelley's exhaustive survey of the theoretical and empirical literature on saving, while providing little positive support, constitutes no barrier to acceptance of Cornwall's conclusion.

Keynes and Hansen are often portrayed as pessimistic about the ability of a developed capitalist economy to operate successfully if population should stop growing. Actually, that was not the import of their argument. They emphasized the possibility that declining population growth might lead to a deficiency of demand *only* in order to strengthen the case for an active fiscal and monetary policy. Given such a policy, the implications of their thesis are, on the contrary, highly optimistic, since any easing of the pressure of private demand on resources opens up more room for the satisfaction of social and environmental needs. This was less obvious in the 1930's than it is now, but even then economists who were both Keynesians and liberals were beginning to feel that increased government spending was desirable for its own sake. Although no one had expressed the idea with the force and clarity of Galbraith's *Affluent Society*, there was a growing feeling that the public sector had been neglected and that it was high time to redress the balance.

Fortunately, as far as policy is concerned it does not really matter whether we can determine the impact of population growth on demand or not. Fiscal and monetary policy are certainly powerful enough to compensate for any deficiency of demand, whether sporadic or chronic, generated by the private sector of the economy. The possibility of such a deficiency seemed very real in the 1930's and to some extent again in the late 1950's and early 1960's. Now with a greatly enlarged nonmilitary public sector and the certainty that major environmental and resource problems are going to be with us for a long time to come, the danger of a deficiency of demand in the foreseeable future seems remote.

Worry about maintaining full employment has given way in recent years to concern lest zero (or slow) population growth cause a labor shortage which would have seriously detrimental effects on the economy. Both Japan and Western Europe have experienced an acute scarcity of labor in the last ten years or so. They have also had slow population and labor force growth. What is more natural than to put the two together and conclude that the latter is the cause of the former?

The first point that must be made is that a labor shortage is never a function of supply alone. This is commonly overlooked in popular discussion in which the existence of a fixed need for labor is taken for granted. Often the need is estimated simply by projecting present trends into the future. That demand is also a variable need hardly be stressed in a gathering of economists.

It follows that a shortage of labor can always be alleviated through fiscal and monetary measures to curtail demand. Most of the participants in professional discussions, such as those conducted by the Organization for Economic Cooperation and Development, admit this. They reject this solution, however, because it would entail reducing employment and output below their existing very high levels. In other words, they accept the continuance of tight labor markets as a desirable goal of policy, though this is rarely stated explicitly.

Thus we find ourselves in a paradoxical situation: Western Europe and Japan are urged to encourage immigration, or more rapid population growth, because they are suffering from a severe shortage of labor. Yet no real relief is offered since the advocates of more rapid labor force growth also favor maintaining heavy pressure of demand on the economy in order to maximize output and employment. What then are the reasons why rapid growth of the labor supply is advocated? Surprisingly, they are related to a desire to raise the rate of growth of output rather than to alter the ease or tightness of the labor market.

Since most of the discussion has focused on immigration rather than on population growth as such, it is necessary to point out some ways in which the effects of the two differ. In particular, there are three commonly cited advantages of immigration which do not apply to growth of the native population: (1) Immigration serves as a "conjunctural shock absorber" (*Kunjuncturpuffer*). (2) Immigration frees domestic workers from the "dirty" jobs. (3) In the early stages of immigration, the newly arrived workers, being without families and not yet having become accustomed to the consumption standards of the host country, typically produce more than they consume.

What these three points add up to is that immigrants can be exploited, at least until they become integrated into the domestic society. These advantages tend to disappear as immigration "matures," i.e., the immigrant workers acquire or are joined by their families and as they obtain the same rights and privileges as the native population. When this stage is reached, the effects of immigration are much like those of population growth. Immigrants become heavy consumers of durable goods, both those produced by the public as well as the private sector. While not going quite as far as Hansen and Cornwall in assessing the effects of population growth on aggregate demand, advocates of labor force growth through immigration do admit that the result is at best a stand-off. W. H. Böhning, for instance, says that the expansion of consumption tends "to nullify the desired satisfaction of labour demand and each foreign worker induces the demand for another one. In this sense economic migration is truly self-feeding . . . " (p. 13).

Again, we are tempted to ask, what is the point of encouraging immigration or population growth, especially in countries already as densely populated and experiencing as serious environmental problems as those of Western Europe and Japan?

The first argument of the proponents is that it obviates the necessity for the structural changes in the economy which would have to accompany an increase (or more rapid increase) in the capital-labor ratio. But is avoiding structural change desirable? While many European writers seem to think it is, some disagree. Ingvar Svennilson, for instance, says:

> An increase in immigration would clearly make it possible to slow down the trend toward more capital-intensive production. . . . Such an argument for a general increase in immigration is, in my opinion, defeatist. The trend we have experienced in certain industries, for example, in the distributive trades, shows that it is possible for production rapidly to develop in a more capital-intensive direction. This has taken place in areas where, under conditions of competition, the free formation of prices has been allowed to have its own ruthless but beneficial way. [p. 42].

Advocates of labor force growth also introduce dynamic elements as a sort of *deus ex machina* to help them break the bonds of neoclassical theory. They attach particular importance to the increased mobility of the labor force with immigration. Bönning, for instance, says: " . . . the smooth growth of the economy under labour import compared with friction in a closed labour market ensures a higher real growth of wages in *per capita* terms with the beginning of the third stage of maturity" (p. 45). Other dynamic factors cited are economies of scale and the favorable influence of labor force growth on technological progress.

Since it is difficult or impossible to provide a refined test for such assertions, we are forced to fall back on crude comparison of rates of productivity increase and labor force growth in different countries. Charles Kindleberger claims to have demonstrated a positive relation between the two for Europe in the period 1959 to 1963. The evidence is not very impressive. It is even less impressive for the period since 1963. Table 1 shows productivity and labor force growth rates in the period 1963–70 for the eight countries Kindleberger carries up to 1963. There is clearly no systematic relation whatever. In the two countries with the fastest growth of productivity, Austria and Italy, the labor force *decreased.* Germany had a high rate of productivity increase with virtually no change in the size of its labor force. There

TABLE 1—LABOR FORCE AND PRODUCTIVITY FOR SELECTED COUNTRIES, 1959–1970

	1959–63		1963–70	
	Labor Force	Labor Productivity	Labor Force	Labor Productivity
Austria	−0.1	5.2	−1.3	6.6
Belgium	0.5	3.3	0.7	4.2
France	0.7	5.1	0.9	4.7
Germany	0.8	4.6	0.2	4.9
Italy	0.1	6.1	−0.4	5.7
Netherlands	1.4	3.7	1.0	5.0
Sweden	1.2	2.6	0.6	3.7
United Kingdom	0.9	1.7	−0.1	2.9

Sources: 1959–63: Kindleberger, p. 27. 1963–70: *U.N. Statistical Yearbook, 1972; OECD Labour Force Statistics 1960–71.*

was, moreover, a general acceleration from the early to the late 1960's of the rate of productivity increase along with a slowing down of the growth of the labor force.

Long-run comparisons are no more favorable to the Böhning-Kindleberger thesis. It is particularly interesting to note that by 1971 Sweden had achieved the highest level of per capita real income in Europe. Sweden neither has a big population nor has it had a high rate of population growth. Perhaps there is at least symbolic significance in the fact that Svennilson offers his defense of neoclassical capital-intensive development against the backdrop of Swedish achievement.

Finally, population growth can alter the age distribution, which can in turn have various effects. We have space here to consider only one of these, namely, the effect of zero, or slow, growth on creativity. Many people fear that a stationary population would, in the words of a French demographer, be a "population of old people ruminating over old ideas in old houses."

Is this fear well founded? As a first step in providing an empirical test, we have made a study of the ages at which Nobel Prize winners in physics did their prize-winning work. The resulting age distribution—which is Gaussian with a mean of 36.2 and a standard deviation of 7.6 years —provides a basis for estimating the probability that a person of given age will do work of Nobel Prize quality. We also determined the age distributions, using United States mortality tables for 1967, of hypothetical populations growing at different rates from −1.0 percent to 3.0 percent per year. It turns out that the relative per capita probability of Nobel-quality work in physics has a broad maximum from zero to 1.5 percent per year growth but falls off rapidly for growth rates less than −0.5 percent and more than 2.0 percent per year. Clearly there is no justification here for fears that slow or zero population growth would have a stultifying effect on creative scientific activity.

One reason for the common misapprehension on this score is the failure to realize that a growing population is "young" primarily because so many of its members are under twenty. The proportion in the crucial age groups—crucial as indicated by Nobel Prize winners in physics—is relatively insensitive to changes in growth rates from zero to 1.5 percent per year. Of course, we need studies of other groups: artists, writers, businessmen, politicians. There is a strong presumption, however, that youth is more important in physics than in most other fields. The average ages of Nobel Prize-winning work in physics (36), chemistry (38), and medicine and physiology (41) indicate a progression from younger to older as the research becomes more applied. Similarly, among the physicists the experimentalists are five years older than their theoretical colleagues. It seems likely that theoretical brilliance is less and empirical experience more of a factor in the success of businessmen and politicians than of scientists. There are some areas, notably athletics, in which youth has a clear advantage. But over the whole range of creative activities fears that a stationary population will bring stagnation seem unjustified.

REFERENCES

G. Ackley, *Macroeconomic Theory,* New York 1961.

W. R. Böhning, "The Economic Effects of the Employment of Foreign Workers," *OECD Working Document,* Feb. 1973.

J. Cornwall, *Growth and Stability in a Mature Economy,* New York 1972.

J. K. Galbraith, *Affluent Society,* New York 1969.

A. C. Kelley, "Demographic Changes and

INTERNATIONAL ENERGY SUPPLY

The Political Economy of World Petroleum

By JAMES W. MCKIE*

Unless we are to confine our discourse to solar energy in the twenty-first century, almost any statement about energy will run a high risk of being overtaken by events. The risk is particularly high for international oil—recently we have not only been overtaken by events, but trampled flat by them. One must of course avoid the temptation to limit one's comments to current developments; *The New York Times* does that better than any of us. This paper looks beyond the present crisis to consider what we can expect from the world oil economy during the next fifteen years or so.

World oil is a mixture of economics and politics. Economics is generally dominant, but not always. The economic model usually cited is the international cartel. The prevailing view up to 1970, which was certainly justified by events, was that the cartel of producing countries—Organization of Petroleum Exporting Countries (*OPEC*)—was ineffective, while the group of oligopolists that produced, transported, refined and marketed the oil in international trade was able to maintain a fairly good defensive position in maintaining its own profits and collecting taxes from consumers. Rapid growth of supply in various parts of the world, especially the Middle East, had caused the world price of crude oil to fall considerably from its level in the middle 1950's, which made the producing countries look for a way of stabilizing it.

OPEC suffered from the centrifugal forces inherent in any cartel struggling with overproduction and persistent excess capacity. The producing countries varied tremendously in resource endowment, need for revenues, and outlook for development. Some were more favorably located with reference to the market than others, yet their collection of location rents provoked dissension. An optimum economic solution for the system required considerably higher reserve-to-production ratios in some countries than in others, again eliciting resentment and suspicion within the group. A few countries had much higher production costs than the general run. There was no political mechanism, no common tradition, no habit of cooperation even within the Arab bloc. The producers were trying to maintain prices much higher than marginal costs, and to enlarge that margin, with no mechanisms in hand for limiting output, allocating markets, disciplining price-cutters, or controlling entry. The outcome was not surprising. Throughout the 1960's, *OPEC* leadership must have wished for something like the powers of the Texas Railroad Commission and the Louisiana Conservation Commission to enable it to impose some form of international production controls.

The situation changed rather abruptly in 1970–71, when rapid demand growth finally overtook capacity—an event punctuated by the cutbacks in Libya and the disappearance of spare production capacity in the United States.

---

* The University of Texas at Austin.

Two conflicting expectations have emerged. The optimistic one (from the standpoint of consumers) looks to the evident huge discrepancy between the price in the Middle East (climbing through $7 per barrel as this is written) and the long-run marginal cost there of fifteen to twenty cents. It expects that discrepancy to result in an irresistible expansion of production from reserves which add up to many times the anticipated demand. There is no practical limit to such expansion, which means that surpluses will inevitably emerge, producing countries will begin jockeying for position in world markets, *OPEC* will break up in the classic manner, and prices in the long run will fall.

That is what our equilibrium models of a competitive industry might lead us to expect, and those who adhere to this view think that competition in a large degree is bound to appear if only the consuming countries do not inflict monopoly on themselves by capitulating in advance to every whim of *OPEC*. They believe that competition is the more likely to appear as the international oil companies lose their position as producers, since producers find it easier to stabilize prices on the basis of uniform taxes paid to producing-country governments. The national companies that are replacing the internationals as producers have no equivalent common basis for price in the absence of central price coordination by the cartel. (Cf. Morris Adelman.)

The pessimistic view does not look at a distant long-run equilibrium, nor at the economic factors only. What counts most is the rate of demand growth in relation to the rate of supply growth. The effective supply is a short-run curve moving through time. It depends in part on the rate at which investment can increase capacity and deliverability in response to high rates of demand growth. In these circumstances the short-run supply in the vicinity of effective demand can remain continuously above long-run marginal cost for a long time—perhaps indefinitely. Cartel members can wait for expanding demand to overtake any excess capacity that may momentarily appear rather than cutting price and aggressively seeking markets. The sellers, in other words, react to expectations of demand growth in the near and middle future as well as to the present situation. A large upward shift in the expected cumulated total of world demand over the next ten or fifteen years has contributed to the recent successes of *OPEC*. (See J. E. Akins and James McKie.)

To account for the picture that has emerged we must also pay some attention to motivations. Supply measures the willingness of suppliers to sell, which depends on other things besides the motive of the individual supplier to maximize his short-run gains. Aside from the recent almost total short-term politicization of the market, over the long run it will be affected by the national policies of the producing countries, which are mixtures of politics and economics. The pressures of competition in the market will not prevent individual countries from exercising considerable discretionary control over their oil policies. The best-publicized aspect of this discretionary control in the Middle East is the decision of certain governments—notably Saudi Arabia and Libya—to hold back the rate of production, and its growth over the long run, below the levels that the market (under the stewardship of the international oil companies) has decreed to be optimal.

These decisions to reduce the rate of expansion of output are not purely political decisions. The governments of Saudi Arabia, Libya, and some other desert countries have a lower rate of time preference for income than do the intermediaries and consumers; they are unable to adjust this rate to the market rate of discount by any convenient means other than conserving their oil resources. They do not expect that other producing areas can or will fill the market gap by increasing the rate of output in these other areas, thus keeping prices down. They cannot find opportunities for domestic investment that will cause the value of the nation's capital, as their rulers evaluate the need for development capital, to increase as rapidly as the oil itself if left in the ground. They are disinclined to increase without limit their holdings of depreciable currencies. All of these are quite rational economic grounds for doing what they propose to do, given the fundamental political attitudes toward economic development in those particular countries.

It is a political fact, of course, that these countries are ringed by others less ascetic, with much more pressing wants for economic development and growth in income, with no comparable source for fulfillment of those wants, and incidentally much better armed, who are beginning to look rather thirstily at the oil-rich power vacuum in those regions of the Middle East. It could make a difference to the rate of price increase if those power vacuums were filled with power having a different political outlook, but not much difference to the direction of price changes as long as world demand kept growing at its projected rates. *OPEC* would still remain strong, and prices would continue to rise. Individual producing countries might temporarily break away from a politically inspired embargo, as Iraq has done recently. But that would not create a condition of surplus supply in the market, and no national producer would find it necessary to cut prices in order to gain markets over the long run. (The producing countries have found during the most recent cutbacks that their incomes went up notwithstanding reductions in output; they are not likely to forget what they have learned about the short-run inelasticity of demand in the face of concerted reductions of supply. Only a demonstration of higher elasticities in the long run is likely to have any effect.)

The most prominent fact in long-run supply, as we see matters now, is the prospective dominance of one country in international petroleum: Saudi Arabia. It is not only expected to have the largest share of world oil production in 1985, but to account for a very large fraction of incremental supply between now and then—all the larger fraction if it abandons presently planned restraints on the rate of expansion. It will be the "swing" producer if it chooses to exercise that role; its decisions will dominate the growth pattern. No longer will the international market be subject to the unmanageable cartel of the 1960's. At long last, the *OPEC* countries may have found their Texas Railroad Commission.

The larger producing countries in the Middle East, including Iran besides Saudi Arabia, will doubtless accelerate their forward integration into refining, transportation, and marketing. The appearance of Flying Camel gasoline in the *U.S.* market (the brand name I owe to Charles Issawi) may disturb our sense of the fitness of things but it will at least give us something to expropriate. Will forward integration by national companies in a parallel network to that of the international oil companies lead to fresh outbreaks of competition and declining prices? Again, the answer depends on whether one thinks that the internationals maintain prices artificially in the face of an inherent surplus of crude oil on world markets, or whether a persistent tendency for demand to run ahead of supply will manifest itself for the next couple of decades, in which case the presence or absence of the internationals will cease to have much bearing on the result.

If the internationals become at best merely service companies for the production and distribution activities of the producing countries, what is left to protect the consuming-country populations if competition among producing countries, or their national companies, fails to emerge or to constrain *OPEC* policies?

The governments of major importers are going to become involved far more than in the past; they will have to have an active petroleum policy within a strategic energy policy. The effect of the current embargo will certainly be a more positive security policy. The options for the United States are wider than those open to Western Europe and Japan, but all can embark on a storage program to give them a short-run cushion against future interruptions and try to develop substitutes and alternative sources. A stored reserve would be useful against an economic embargo as well as a political one, but it has one drawback: the opponents know just about how long it will last.

Consuming countries, especially Common Market countries and Japan, are also likely to try out some new ways of bargaining. We can expect them to do several things; some have already begun:

1. They may set up national companies of their own, or form officially sponsored trading bodies, to centralize the purchase or other acquisition of oil from producing countries by various means (including purchases from international companies or allocation of part of the final market to them.)

2. Some importing countries will try to conclude bilateral agreements with the governments of certain producing countries (as France has already done) to assure themselves of a preferred if not secure position as buyer of that country's output. Unfortunately, this form of vertical quasi-integration may guarantee mainly the right to pay more than the world price for an assured supply.

3. Some governments will attempt to form a bloc—perhaps an Atlantic-American-Japanese Community—to coordinate policies on storage and stockpiling, emergency rationing, bargaining with producers, development of alternative sources, etc. (See the suggestion by Walter Levy.) At the extreme, this could become a monopsony with sole powers to buy from *OPEC* countries. The analogy is tempting. There are two problems with this approach. The first is a well-known property of bilateral monopoly: it often leads to stalemates in bargaining, which can only be resolved by a "strike"—in other words, it makes producer embargoes or consumer boycotts more, not less, likely. The second is that the buyers would not really form a monopsony but an oligopsony with some of the attributes of a cartel. In a tight market, with the respective national populations clamoring for maintenance of accustomed consumption and rapid growth, the temptation to break away and pursue one's own benefit would be very strong. The consuming countries would, in other words, be subject to the same kinds of centrifugal forces that nullified *OPEC* during the 1960's. The behavior of the European buyers during the recent days of the Arab embargo does not encourage the belief that coordination of consuming countries' policies would be at all easy.

The rate of growth of demand is obviously a key question in the long-run behavior of international petroleum markets. For illustration, let us contrast the effects of a rate of growth of 8 percent (roughly the average rate of increase in world consumption for oil during the last five years) and 4 percent per annum in world demand between now and 1985. Let us consider only the (present) non-Communist world. Proved and published crude oil reserves (excluding the Communist bloc) at the end of 1972, according to the *Oil and Gas Journal*, were 77.5 billion tons. Consumption in 1972 was about 2.2 billion tons: a comfortable average reserve/production ratio of 35 to 1, though it varied considerably from region to region. (All these figures are from *BP Statistical Review.*) If consumption grew at 8 percent per annum, it would mop up about 51 billion tons of the present 77.5 billion by 1985. How much in new reserves would be added by 1985? Let us assume something over 50 percent of the recent figure, or another 40 billion tons or so outside Russia, China and Eastern Europe. This is perhaps less than the world industry *could* add in that space of time, but a 50 percent replacement of reserves in thirteen years would be no mean achievement.

Proved reserves in the "free world" (77.5 billion at the end of 1972) amounted to about one-third of the estimated ultimately recoverable supplies of oil. (These are estimated by Shell Oil Company at about 1.8 trillion barrels, or 240 billion tons. See also the estimates reported in *U.S. Energy Outlook, 1971:* at usually assumed recoverability ratios, upwards of 2.1 trillion barrels, 290 billion tons, recoverable for the world, or less than 250 billion tons for the non-Communist world.) An increase of 40 billion by 1985 would bring the developed fraction almost to 50 percent of the recoverable total.

Meanwhile, however, consumption would increase from 2.2 billion to almost 6 billion tons. Reserves used in the interim (51 billion tons) would not be fully replaced; reserves in 1985 would be about 66 billion tons, and the reserve/production ratio would have fallen to eleven to one. *If* such a process were anticipated to begin in 1973 and actually materialized, there wouldn't be much price-cutting by *OPEC*.

On the other hand, assume that demand will grow at 4 percent per year. (This is a higher rate than was being assumed in many authoritative publications a few years ago, and higher than that projected in *The Oil Import Question.*) The total consumed, 1973–1985, would be about 38 billion tons, less than the assumed new discoveries of 40 billion, and the remaining reserves of 80 billion in 1985 would be over twenty-one times the production of 3.7 billion in that year. But if lower rates of demand growth were anticipated, expected world prices would not increase as fast as in the preceding case; reserve additions would be somewhat less as a result, but the ratio to production would almost certainly be higher in the 1980's than in the 8 percent case.

From the observation point of late 1973, we do not know with any certainty what discoveries and development will occur over the next decade and a half (nor, of course, will they cease in 1985; we sometimes speak as if getting through to that target date comprised the whole problem.) The above assumption of a 50 percent increment could be laughably wrong. The high growth of demand, resulting in higher prices, could stimulate larger additions. But what we know *now* indicates that over two-thirds of that increment would probably be developed in the Middle East, and over half would be in Saudi Arabia. Hence the incremental demand must be concentrated on that one supplier, which will lead to difficulties if demand grows faster than the Saudis are willing to expand output; as mentioned previously, Saudi production is not necessarily going to rise to a ratio of reserves that Western oil companies would consider efficient. If they hold production to 8 or 10 million barrels per day instead of permitting planned increases to 20 million within the next few years, there is no evident place for the Western world to turn to make up all of that difference, though some extra output may be forthcoming from Iran and Iraq.

Nor do we know with certainty what demand will be during the next 15 years; that will depend in part on the energy policies that Western countries and Japan decide to adopt. A high (and climbing) world price of oil will hold down demand growth, to be sure, but at the cost of enormous scarcity-rent transfers to the producing countries. Availability of substitute fuels at certain price horizons will help, of course, by increasing the elasticity of demand when those points are reached. If high demand growth is strong enough to sustain high prices, the consuming countries' governments can transfer some of those rents back to themselves by imposing large import taxes on foreign petroleum. But it may well be doubted whether a projected growth rate or shift of the short-run world demand curve for petroleum of 8 percent per annum would be consistent for long with the price levels that the key producing countries are now contemplating. Moreover, the consuming nations can help to kick the demand off the highly inelastic short-run curve onto, or toward, the more elastic long-run curve by adopting specific supplementary policies to retard the growth of demand and encourage the use of substitutes. The changes in patterns of use should not be regarded as adjustments to a transitory situation.

Will these developments—appearance of higher long-run elasticities of demand at higher prices, supplemented by efforts of consuming countries to restrain demand growth still further—cause a long-run surplus of supply to appear again so that unrestrained rivalry will break out and prices will decline as they did in the 1960's? I doubt it, though any forecasts could be totally destroyed by discovery of four or five new Kuwaits off the coasts of Australia, Massachusetts, Alabama, Ireland, and Norway. I think that restraint of demand growth at best will curb exploitation of the Western world (both economically and politically) and help to remedy some deficiencies in the economic calculus of conservation.

It is to be hoped that the world has begun to learn what exponential growth rates of 6 or 8 percent for petroleum can mean in the long run. Even 3 or 4 percent could be too much for an exhaustible resource unless suitable substitutes can be developed in time, including those still beyond our technological horizons. Frantic efforts on the supply side to develop additional reserves of petroleum really only shift the exhaustion curve toward the present. In view of the great uncertainties of technological development, one may legitimately doubt whether we can buy enough time by working on supply side alone.

Along with the economic uncertainties of growing dependence on petroleum we must not forget the political ones. Increased dependence means increased vulnerability in both dimensions, and the political may well be more significant than the economic. What might have been said about the political effects as late as September is no longer necessary: they are unrolling before our eyes. We will probably go to considerable lengths to prevent them from happening again. Reduction in projected demands for imported petroleum would help the Western Alliance to retain some measure of freedom of action in world politics.

REFERENCES

M. A. Adelman, "Is the Oil Shortage Real?" *Foreign Policy*, Winter 1972, *9*, 69–107.

J. E. Akins, "Remarks to the Independent Petroleum Association of America," New Orleans, May 9, 1972.

W. J. Levy, "An Atlantic-Japanese Energy Policy," Statement presented before the Europe-America Conference, Amsterdam, March 1973.

J. W. McKie, "Balancing the Demand and Supply of Oil," in J. Schanz, ed., *Balancing Supply and Demand for Energy in the United States*, Denver 1972, 73–90.

British Petroleum Company, *BP Statistical Review of the World Oil Industry*, London 1972.

National Petroleum Council, *U.S. Energy Outlook: An Initial Appraisal, 1971–1985*, Washington 1971, *2*, 41–53.

Cabinet Task Force, *The Oil Import Question*, Report of the Cabinet Task Force on Oil Import Control, Washington 1970.

Shell Oil Company, *The National Energy Position*, Houston 1972.

Politics, Economics, and World Oil

By M. A. ADELMAN*

Zen Buddhist monks used to torment novices by asking: "What is the sound of one hand clapping?" That sound has become deafening in recent years: the official predictions that because world oil consumption will increase, oil must grow more scarce and the price must increase. But scarcity is the pressure of *demand* upon *supply*. To omit either element is nonsense. They are united in the true measure of scarcity, long-run marginal cost. The only relevant question is whether, as consumption grows, society must keep putting more or less into the ground to get out another barrel.

Long-run marginal cost is mostly the return on the investment needed to develop additional capacity. Failure to discover new flush reservoirs means ever more intensive development of old fields, hence rising development investment and cost per unit, and rising prices. Anticipated price-cost increases delay development of some deposits, forcing more intensive work on the remaining ones, hence higher costs. Thereby a development cost increase serves as a distant early warning signal of future scarcity, bringing it into the present. Conversely, a stable (or declining) marginal cost means no greater scarcity, and this is the actual case. For the Persian Gulf, or even for the whole world outside North America, real costs have been sharply declining, and even if they were now to reverse course and climb, as is always possible, they would be a negligible fraction of price. The current flood of projections from here to eternity is a pitiful and futile attempt to replace the price-cost thermometer-thermostat, but they are official truth. One projects "needs" and "amounts available" to find a "surplus" or "deficit" regardless of elasticities of demand or supply.[1]

A Royal Dutch Shell executive sums up the world oil market: "The underlying situation of supply and demand remains one of potential surplus. Yet the producing countries manage to reap the rewards of a sellers' market by creating a producer's monopoly" (Geoffrey Chandler).

I. Monopoly

The multinational oil companies are not junior partners but rather agents of that monopoly, the members of the Organization of Petroleum Exporting Countries (*OPEC*) (but not *OPEC* itself). Aside from short-term flights, the price is now around $8 and (probably) close to the long-run profit-maximizing level set by the competition of substitute energy sources, as the Shah of Iran has stated (*New York Times* [*NYT*], Dec. 24, 1973). The official truth stated by Secretary of State Henry Kissinger, that prices have risen because of a surge of demand against inelastic supply (*NYT*, Dec. 12, 1973), is in utter conflict with the fact of enormous supply elasticity at cost of at most one-fortieth of the current price.

The popular slogan "avoid overbidding" suggests that oil prices have been bid up by demand exceeding supply, which is untrue, and also betrays a misunderstanding

---

* Massachusetts Institute of Technology. I wish to thank Harry J. Colish, Richard L. Gordon, Richard B. Mancke, and Joseph L. Yager for comments on an earlier draft; errors are my own.

[1] Elementary economics is ignored in grain as in oil: the Department of Agriculture's Economic Research Service was never consulted on the notorious 1972 wheat sale to Soviet Russia (*NYT*, Oct. 7, 1973).

of what has been happening in 1972–73. Current supply-demand fluctuated, with occasional excess capacity. But the demand for crude for *later* delivery was insatiable because buyers knew prices were going to be raised. Buyers had little downside risk. If the producing countries delivered at the contract price, buyers would make a speculative gain; if they delivered at the expected higher prices little would be lost. Whereupon the *OPEC* countries turned around and cited the rising contract prices as a reason for raising their taxes—thereby putting a firm tax floor under the higher prices and validating the expectations. "Reasons" are as plentiful as blackberries; what matters is the power to raise the price of oil close to the cost of (expensive) substitutes.

Monopoly means control of supply, hence power to stop it, hence dependence and insecurity. Food is more essential than fuel, yet nobody is "dependent" on any farmer or on all farmers together, because farmers cannot act together to control and if need be withhold their production. Our "dependence" on imports exists only because of the cartel and (in the short run) the Arab majority bloc. Its history is extremely important: "those who ignore the past are condemned to repeat it," and we have already repeated it once. The key words in that history are *threats* by the producers, and *collaboration* by the consuming nations, especially the United States. The threat of an embargo gave the cartel its first triumph: the Tehran "agreements" of February 1971, whose expected and actual effect was to raise prices at a time of slack demand. The documented record of the 1970–71 events shows that only *after* American-sponsored capitulation to producing country demands in January did anyone dare voice public threats. The American policy maker did not blush to tell a Senate committee that threats had been made *before* the capitulation—and ceased upon his request; see M. A. Adelman (1973). He later explained that the threats had been made privately (James Akins). This evades the issue; threats are made in private so that they may be denied, reinterpreted, or repudiated. And to say that the threats ceased is completely false in the light of the numerous public statements which culminated in a formal *OPEC* resolution, issued just before the Tehran agreements, threatening "total embargo"—and equally numerous threats since.

The first triumph of blackmail announced more to come—as some were then denounced for saying. Our government "expected the previously turbulent world oil situation to calm down following the new agreement." In fact, the five-year Tehran agreement lasted four months, and after several "revisions" was pronounced "dead" last fall when the Persian Gulf nations unilaterally raised prices. Perhaps they were bored with what an oilman called "the charade of negotiations." (*NYT*, Oct. 19, 1973.) But the American policy maker may be right in claiming that the Tehran agreements "worked well" (Akins)—from his point of view. So also with the proposal for preferential entry into the United States for Saudi Arabian oil—the most insecure source conceivable. The State Department was "enthusiastic" (*Oil and Gas Journal* [*OGJ*], Oct. 9, 1972), for reasons not explained. Nobody can doubt that its "exaggerated talk of an energy crisis greatly strengthened the bargaining power of the Arab states" (Petroleum Press Service [PPS] Nov. 1973).

II. Shooting War and Economic War

Middle East politics, specifically the Arab-Israel tension, have had no effect on the price, and a Middle East settlement will do nothing at all to keep the price from increasing to the monopoly level. The producing nations will take what they can get. The monopoly revenues make peace unlikely. (See below.)

The shooting war and economic war waged by a subgroup of the cartel—the Arab oil producers—were invited by repeated American statements, of which the public record is probably only the tip of the iceberg, and whose complete exploration would richly repay a Congressional inquiry. The Saudis were told they were the last best hope of civilization, we had to have their oil, and would they not please produce it, even though it was not (we said) in their economic interest to do so? No revenues were high enough to induce the Saudi government to agree to big production increases; something extra must be done for them. (*Wall Street Journal* [*WSJ*], Aug. 15, 1973; *OGJ*, Sept. 10, 1973.) This was a self-fulfilling prophecy. For if we believe it and are willing to do something extra for the Saudis, they are glad to demand it.

The buyer is asking to be had who tells a seller, "I know you don't want any more business but please just to do me a favor won't you sell me something?" There are few such buyers because they don't stay in business very long. Not so in government.

Sheik Yamani, the Saudi Arabian petroleum minister, recently asserted that before the recent cutbacks, when Saudi Arabia was producing eight million barrels per day (*MBD*), "we were producing at a much higher rate than what we should for our economy (*Meet the Press* [*MP*], Dec. 9, 1973). And that was a sacrifice on our part." It amounted to "losing money." Sheik Yamani says, without a smile, that his government has been producing not for its benefit but for sweet charity. He speaks as a man who expects to be believed. But that's no wonder, for the United States Government was saying this publicly before he was.

But if eight *MBD* is a production rate "much" too high for Saudi Arabia's good, twenty *MBD* is catastrophically too high, and we will owe them three times as much or more in 1980 than we do now. To keep the oil flowing, we will impose a just peace in the Middle East this year. Next year it will have to be even more just, and the year after that . . . and so on.

This official truth about needing to do something for Saudi Arabia, because it is not worth their while to expand output, is all implicit, never set down for analysis. But it appears to rest on two assumptions. (a) "Oil in the ground appreciates faster than money in the bank," abbreviated *OGMB*. It is often embellished by saying that the dollar has depreciated, and prices on the New York Stock Exchange have gone to pot, ergo, there is no place to invest. In fact the dollar may be an undervalued currency, or payment may be in another, and in any case the annual volume of capital formation in the developed world (not to mention the total stock of existing purchasable assets) is many times the future revenues of even Saudi Arabia. But let that go: *OGMB* is at best meaningless without specific numbers. The current price of Persian Gulf oil is about $8. If 8 percent is a safe interest rate, then a barrel is worth holding, instead of a corporate bond, for four years for an expected price rise to $11, and for nine years for an expected price of $16. The price is not going to appreciate indefinitely at 8 percent per year: it is not going to $32 in eighteen years, nor to $64 in twenty-seven years. But Saudi Arabian crude reserves are fifty times current output, and can be greatly increased at negligible cost. They are held back in order not to wreck prices. *OGMB* is an irrelevance in a noncompetitive market.

(b) Saudi Arabia (and others) will limit their oil revenues to what they "need." This means—if it means anything—that they will hold back output short of the monopoly optimum, i.e., the point where it maximizes the present worth of their assets. It is an odd assumption, to say the least, and quite unsupported. If King Faisal acts like a true dynast to serve his successors, family, retainers, friends, etc., the best way to insure this is to maximize present worth.

We are better off with less talk of "need" and a little thought about economics. Saudi Arabia, like the U.S. Steel Corporation or the Texas Railroad Commission in other days, has the usual problem of Mr. Big in a cartel: find the combination of price and quantity which will maximize group profits—or more generally, best serve the economic interests of the producers. They can fix the price and let the price determine quantity; or fix the quantity and let it determine price, but these are only two routes to the same goal. No blandishments will make them expand output; anyone who thinks he can persuade them is merely stroking his ego and reminding us how right was a Scottish professor of moral philosophy who warned against the "overweening conceit" of men "in their own abilities." Repeated assurances of how badly we need them are simply taken as evidence of inelastic demand and signal the monopolist that there is greater profit in even greater restriction.

The drift of American policy was visible in these statements that we owed Saudi Arabia, to whom we sent as ambassador the principal architect and defender of the Tehran "agreements." His earlier statement that "a seller's market arrived in June 1967" disregards three years' price decline but reveals his belief that the Six-Day War was a calamity to be reversed. He "think(s) the *OPEC* countries should be granted substantial increases," in order to induce alternative energy sources needed "to avoid an energy crisis in the 1980's, or 1990's," a "crisis" again assumed, never explained. Also, "price increases will hurt America's commercial competitors Europe and Japan," and Saudi Arabian revenues would mostly be invested in the United States (*Economist* [*Ec.*], Nov. 26, 1973).

Saudi Arabia "planned the Arab strategy for the [1973] Middle East War," both shooting war and economic war (*LeMonde* [*LM*], Oct. 9, 1973; *NYT News of the Week in Review* [*NWR*], Oct. 14, 1973; *OGJ*, Oct. 15, 1973; *NYT*, Nov. 10, 1973; see also *NYT Mag.*, Nov. 18, 1973). King Faisal and Prince Saud al-Faisal had stated they needed to put pressure on the United States. But "the *U.S.* can get along without Arab oil until the end of the decade" (*OGJ*, Sept. 12, 1973). Therefore it was necessary to reduce total production deeply and deprive others of more oil in order to deprive the United States of less. A selective embargo was taken seriously by our principal policy maker but by nobody else (Akins 1973).

III. Surrender Without a Fight

The cutbacks have been a great political success. We are right back to the 1930's, when European nations looked for a deal with the aggressor in the hope he would go jump on somebody else, and when German generals opposed to aggression were discredited by the willingness of the Western powers to give away other people's lands and lives; so too the moderates among the Arabs. For, confronted with the cutbacks, the Europeans and Japanese stood clear of the Americans, however dangerous that was for them. More important, in my opinion, was their inaction at home: oil stocks were not spread over time by rationing, i.e., not used as a defensive weapon to gain time wherein to carry out a plan, but as a means of putting off the unpopular decisions to curtail demand. Most important was European eagerness to collaborate with the Arabs rather than each other. "Arabs don't have to police their own boycotts. Sycophant nations are doing it for them" (*WSJ*, Nov. 6, 1973). The Common Market countries refused to ship or pool oil resources, as requested by the Dutch who had been picked out as a special victim. Apparently the Dutch are getting some covert help—but only after they threatened to cut off natural gas deliveries to France and other nearby countries.

Japan had been more pro-Arab than any large country but France, and had stood aloof from other consuming nations, lest they offend (Petroleum Intelligence Weekly [*PIW*], May 14, 1973), only to find itself accused of "odious neutrality" (*NYT*, Oct. 18, 1973). Saudi Arabia was ready to make new demands "because of their success in recent years in enforcing a boycott . . ." (*WSJ*, Nov. 7, 1973; *NYT*, Nov. 9, 1973).

A cut in British deliveries ". . . is clearly causing embarrassment to the government, which . . . had received assurances [sic] from Arab countries . . . " (Platt's Oilgram [*POPS*], Nov. 20, 1973). The French government is embarrassed over reduced supplies (*NYT*, Nov. 20, 1973). Such governments are especially reluctant to begin rationing because it would be an "admission of failure," i.e., groveling did not insure oil supply (*PIW*, Nov. 19, 1973; *LM*, Nov. 23, 1973).

The servility of consuming governments, playing the Abbe Alberoni to the Arabs' Duc de Vendôme (see Luigi Barzini), has made the original Arab demands of no importance. A weapon which makes consumer nations shake like jelly cannot be contained by a scrap of paper enumerating Israeli security or Palestinians' rights, etc.—there are far bigger objectives now to be considered. Moreover, Saudi money, to be multiplied many fold, can procure more arms from many sources, freeing the Arabs from whatever control the Soviet Union might exercise. This makes fresh wars likely if not inevitable, especially when the Saudis begin shopping for the nuclear weapons they can well afford. Already there is a semiofficial Egyptian call for nuclear weapons, which would cost only an estimated $1 billion, stimulated by "a high-level Washington visitor" to Cairo (*NYT*, Nov. 24, 1973).

IV. Economic Failure, Political Success

Yet the cutback failed badly to reduce American supply. At its maximum (as of December) it amounted to 4.7 *MBD*, about 14 percent of all oil moving in international trade. Hence had there been just enough leakage and diversion to put us as well off as oil importers generally, our import loss would have been about 14 percent. Now, for the four weeks ending November 16, our combined imports of crude and products averaged 6.55 *MBD*. Since the boycott date was October 17, and Persian Gulf—U. S. transit time is about a month, this amount measures the preboycott level of shipments. For the next four weeks, through December 14, the average was 6.10 *MBD*, indicating a loss of about 450 thousand barrels daily, 7 percent of imports, about 2.4 percent of total supply. The truly vulnerable place was the East Coast's heavy reliance on residual fuel oil, much from Caribbean and Canadian refineries which also ran some Arab crude oil and might therefore be forced to stop all shipments to this country in order not to lose some supply. An Arab resolution of November 26 to cut off the Caribbean and other transshipment centers (*OGJ*, Dec. 10, 1973) shows that by mid-November the Arabs realized their failure and their resolve to damage this country where they could. Yet on November 8 our ambassador to Saudi Arabia had warned that the plight of the East Coast would be "critical" if Arab oil supplies were not increased "in a matter of days" (sic) (*NYT*, Nov. 10, 1973). This was wildly untrue.

The Arabs' 25 percent cutback in their production was scheduled originally to keep increasing 5 percent per month until Israel withdrew to her 1967 boundaries and "the legal rights of the Palestinian people" were restored. But the Arab oil exporters' meeting of December 26 reduced the cutback to 15 percent, ignored the two conditions, and let it be understood that the cutback would be cancelled upon Israeli withdrawal from the west bank of the Suez Canal (*NYT*, Nov. 26, 1973). Furthermore, the "friendly" nations (Britain, France) were guaranteed Arab oil even in excess of the base amount (September 1973), which means that their previous imports of non-Arab oil are completely freed for the not-so-friendly (Japan) or the unfriendly nations (United States, Netherlands).

To what extent this failure of the production cutback to reduce *U.S.* supply is due to cheating by the Arab producers and to diversion of non-Arab oil is as yet impossible to say.

V. Monopoly Harmful to Consuming Nations

Relief at this failure should not obscure the fact that the oil cartel is very harmful to American interests. (a) In 1974 customers, including us, will be paying out well over $100 billion, and over 1972–80 cumulative the transfer to the producing countries will be several times that. The richer these nations are, the better they can maintain an embargo to make supply yet more insecure—as the Arab production cutbacks remind us. (b) The world monetary system will be harmed by huge amounts of liquid funds ready to move at a moment's notice, not to serve the holders' malice (the usual straw man) but for self-protection. Controls on capital movements to prevent this danger will in themselves be harmful. (c) Restrictions on American imports, because of the expected oil deficit, have already embroiled us with our main trading partners in Europe and Asia, not only because of John Connally's bluster and bullying but also over the substance of our demand to get more than we give (*NYT*, May 10, 1973).

(d) The risk of mineral exploitation in less-developed countries is much greater; concessions and contracts are now worthless. (e) The hope of a rule of law for the world's oceans has gone by the board because of the hugely inflated artificial value of any possibility of oil. (f) A vast arms buildup is just beginning in the Persian Gulf. Producers have billions available and every little patch of barren ground or barren seawater is actually or potentially worth fighting over.

The arms buildup reminds us that although the oil monopoly will cost us dear, there will be gains for exporters and for contractors for construction, investment management, public relations, etc. There will be plums for many in the industrialized nations and crumbs for less-developed countries. Those "working for the petrodollar," paid or enriched by the monopoly, are highly influential. Moreover, each industrialized nation can hope that the burden will be borne by them in proportion to their oil consumption, but that they will get a disproportionate share of export and investment business. M. Pompidou appears to have talked to King Faisal of little else but French exports during their May 1973 meeting (*LM*, May 15, 1973); he is now "shocked" that anyone thinks exports have much to do with his Middle East policy (*LM*, Nov. 19, 1973).

VI. Implications for Policy

Only in the long run can we get the cartel off our backs, and it will not be easy, quick, or cheap. It is necessary but no longer sufficient to stop the oil producing companies from being the vehicle for the price-fixing agreement of the producing governments. (1) Expelling the companies and losing their know-how would be a huge waste of resources, harmful to all. But if they simply produced (and developed and explored) and were paid in money or a modest share of the oil, the producing countries would have to do their own selling and monitor thousands of transactions all over the world. The companies have managed the difficult task of determining output shares because they have sold the bulk of the final product; the producer nations would inherit the task without the means. Nothing in the history of the trade suggests they would succeed; even the tight cartel of the 1930's was eroded, and it never faced an independent refining industry.

A managing director of Royal Dutch Shell has well said that in buying from producing countries the multinational oil companies "have formidable advantages." (See G. A. Wagner and A. Glimmerveen.) Once they become "formidable" buyers of crude oil rather than tax collecting agents, the market will look considerably different from what it does today. The oil companies are a big gun pointing toward the consuming countries, which ought to be pointed the other way. Hence real nationalization is greatly to the advantage of the consuming countries.

The producing countries may yet oblige us, as did Algeria and Iraq, by first expelling the companies and then inviting them back as contractors or by doing their own selling of most of their oil as "participation." This is good for the individual country in the short run and bad for the group in the longer run—the classic cartel dilemma. It is imprudent to assume they will be so helpful, but the chances of this happening look better in late 1973 than I expected a year earlier; see Adelman (1973). Perhaps such prophecies will be realized as those of Thomas R. Stauffer in 1970: "We conclude . . . that prices will probably sink below the $1.15 level and that . . . non-concessionary oil will drive out concessionary oil." The price-undermining effect of direct sale by the producing countries' national companies ("nonconcessionary oil") remains a key variable if the consuming countries want to make it one.

The American government ought not to force American companies into being contractors, since they would merely be displaced by European or Asian companies. It must be done in unison or not at all.

By the time most consuming nations see their interests a bit more clearly, some will have taken another step: put oil imports under quota, to sell the tickets on sealed competitive bids. Any country which wished to expand or even retain its market in the United States would have to share its gains with the Treasury. This would not reduce the price of oil to the consumer. There would be in effect a tax on imported oil which would keep the domestic price level high also. (In my opinion a high energy price is desirable to reduce pollution and congestion. Those who disagree with this policy judgment may yet prefer to have the money go to the American not the Saudi government.)

If the producing countries succeeded in collusively fixing quota ticket prices, we would be no worse off, but chances of success are small because it would not take much cheating to fill the quota. Detection of cheating would be difficult and might be made impossible by Theodore Moran's suggestion that prices in any given bid be kept permanently secret. There would be no way of knowing whether any country's higher exports were due to its cheating.

The current price level is so much higher than the cost of producing oil, *even in high-cost deposits*—see Adelman (1972)—that trickles, then streams of new supply in the 1970's are a foregone conclusion, and they have been the bane of all cartels. Given supertankers and superports, a barrel of oil anywhere in the world is a barrel everywhere, at a transport cost of a dollar, which is little compared to the producing nation's profit. Only the shortage of men and materials keeps this potential from becoming actual. But even now the producing countries are not deceived about the "world oil shortage." Saudi Arabia, as mentioned earlier, tried for preferential entry into the United States, which only makes sense when more people are trying to enter than there are places to set them. Venezuela keeps proposing worldwide prorationing. Iraq expelled the Iraq Petroleum Corporation from the largest oilfield because they refused to expand output, which under the new regime will have doubled or tripled from 1972 to 1975. When the consuming countries want to get rid of the burden they can; but at present there is no will, hence no way.

This brings us back to the dismal present and decisions to be made soon. The Arabs have failed to cripple us; the Administration is trying to snatch defeat from the jaws of victory to serve some grand design not yet revealed to us. Our greatest immediate danger lies in a super-Tehran agreement for "cooperation" of producing and consuming states, announced by a flourish of trumpets on a TV spectacular, with the same promise made by the same people who brought us the first Tehran that *this time* "the previously turbulent world oil situation" will *really* "quiet down." The ambassador to Saudi Arabia, who in 1972 told the Arabs that it was in their interest to curtail output, told us that for lack of oil our condition would be "desperate" by 1976, (Adelman 1973) and thought the Tehran agreement had worked well, etc., has suggested a world commodity agreement to set oil prices and ensure availability (*NYT*, Apr. 16, 1973). It would be a one-way street, preventing independent action by consumer states to promote price reductions. But if the monopoly holds and the price can be rigged higher, up it will go. A few weeks ago, in proclaiming the Tehran agreements dead, Sheik Yamani supplied a classic formula: "We in Saudi Arabia would have liked to honor and abide by the Tehran agreements, but . . . " circumstances had changed (Middle East Economic Survey (*MEES*) Sept. 7, 1973). Sheik Yamani may one day say that he and his colleagues would have *dearly loved* to honor and abide by the Kissinger agreements, but . . . circumstances, etc. The supersubtle diplomat is no match for the fellow who grabs what is in his reach, then asks if you want to fight to get it back. But it may not even be necessary. For in waving proudly an "understanding" with Saudi Arabia to let output increase to 20 *MBD* or whatever, our government will not realize that there is no meaning whatever to an agreement which does not specify both quantity and price. For if Saudi Arabia's interests are better served by producing less, it raises price to where there is less demanded.

As regards supply outside this country, a sound world oil policy for the short run is to do and say nothing. There are some virtues in necessity. Without a world agreement, each producing nation will seek to maximize its own profit. If Saudi Arabia will for years play the statesman and hold back on output expansion, we are no worse off; if they retaliate against any rivals, we have gained enormously. Similarly with the consuming countries: some of them will recover from their panic and will begin inviting producers to make some special deals for disguised low prices, to put the cartel on the slippery slope. This country needs not *ordnung* but disarray in the cartel. But we cannot by statesmanlike action cure the nonexistent world oil shortage.

However, there are some matters where action may help. Sheik Yamani warned in early 1973 that any attempt at consumers' self-defense meant "war," and "their [i.e., our] industries and civilization would collapse" (Platt's Oilgram News Service [*PONS*], Feb. 22, 1973). By November 9, he and his colleagues "are letting the word out that the present cutbacks in oil output are the limit." The reasons mentioned are possible Western responses: food, manufactures (including armaments), and military action (*NYT*, Nov. 10, 1973). They who had talked of "war" and suited the action to the word understand the language. We had better learn it quickly.

There is as yet no weakening in our infatuation with Saudi Arabia, to whom we seem resolved to return bounty for evil done to us. In late November "a very high official in the Nixon Administration who is a policy maker in this area" told a reporter "he feels King Faisal . . . at the last minute would prevent any serious economic harm from being done to this country because he is at heart a friend of the United States" (*MP*, Nov. 25, 1973). Meanwhile, King Faisal is "angry with Mr. Sadat" of Egypt for being too cooperative with the Americans (*Ec.*, Nov. 24, 1973). Without doubt the Saudis feel they have every right to be hostile to the United States, and it is not for an American to say they are wrong. But our safety demands that we recognize which way is down.

The Saudi connection, which our government values so highly, is no asset but a heavy liability. The profits of Aramco, whose protection is a perfectly legitimate national objective, will be kept at a level needed to secure incremental investment, and can scarcely amount to a billion dollars annually even if Aramco reaches 20 *MBD*. If the Saudi investment portfolio reaches $100 billion, a 0.1 percent per year management fee is the most the management company can reasonably expect. This is less than the extraordinary expenditures already forced upon us this year by King Faisal's shooting war, and it is insignificant compared with the losses of national product here and throughout the world, due to the oil embargo: 1 percent of *GNP* lost is $13 billion per year.

In war one seeks not to be strong everywhere, but only at the strategic points. For the non-Communist world the decisive point is the United States. This country should immediately take steps to separate itself completely from Arab oil sources. Once we are beyond the reach of oil cutoffs, they can no longer pressure us. Then there is no profit in tormenting Europe and Asia, and risking retaliation, as an indirect means of pressuring the United States.

Our overseas imports before the cutback were about six *MBD*. Future imports will for a time be larger, but will come nowhere near the ten or more *MBD* freely predicted a short time ago, because of the drive for greater self-sufficiency. The four largest non-Arab oil exporters—Iran, Venezuela, Nigeria, and Indonesia—already produce thirteen *MBD*, and their production will grow substantially, Iran alone being a good bet for 10 million *MBD* in a few years, especially if we act. (Our current ambassador to Saudi Arabia insisted in September 1972 that Iran *had been* interested in production increases, but no longer (*OGJ*, Sept. 25, 1973), which was contradicted by previous public evidence (*OGJ*, Aug. 14, 1973, Sept. 4, 1973, Sept. 13, 1973), and also the expansion program decided early in 1973.)

Two routes ought to be examined. One is to bar or penalize imports from countries declaring embargoes against us or pressuring third parties to embargo us (see above). Or the United States could make contracts with any countries desiring preferential entry, in return for which they would guarantee certain minimum amounts. We would of course have to promise—and keep our promise—to pay the very high world prices. But as shown earlier, this price will likely be in the neighborhood of what it would cost us anyway to produce at home from substitute sources. Richard Gardner has embarrassed our government by pointing out that Saudi Arabia has violated their treaty with us providing for mutual most-favored-nation treatment (*NYT*, Dec. 19, 1973). We need only tell the Saudis their embargo on shipments to us is henceforth permanent, their status having been cancelled by their own act.

As George F. Kennan, a respected scholar and ex-diplomat, has well shown, in saving ourselves, we save our friends abroad, by making boycotts against them unrewarding and therefore unlikely.

REFERENCES

M. A. Adelman, *The World Petroleum Market*, Baltimore 1972.

———, "Is the Oil Shortage Real? Oil Companies as *OPEC* Tax Collectors," *Foreign Policy*, *9*, Jan. 1973.

J. E. Akins, "This Time the Wolf is Here," *Foreign Affairs*, *51*, Apr. 1973.

L. Barzini, *The Italians*, 1964.

G. Chandler, "Some Current Thoughts on the Oil Industry," *Petroleum Rev.*, Presidential address, Institute of Petroleum, *27*, Jan. 1973.

G. F. Kennan, "And Thank You Very Much," *New York Times*, Op. Ed., Dec. 2, 1973.

T. R. Stauffer, "Price Formation in the Eastern Hemisphere: Concessionary versus Non-Concessionary Oil," in Zuhayr Mikdashi, et al., *Continuity and Change in the World Oil Industry*, Beirut 1970.

G. A. Wagner and A. Glimmerveen, Presentation to a meeting of financial analysts in The Hague on Oct. 4, 1973.

American Petroleum Institute, *Weekly Statist. Bull.* (*API*).

British Petroleum Statistical Review of the World Oil Industry (*BP*).

Bureau of Mines, monthly releases (*BP*).

Economist, London (*Ec.*)

LeMonde (*LM*).

Meet the Press, NBC Television (*MP*).

Middle East Economic Survey (*MEES*).

New York Times (*NYT*).

New York Times, News of the Week in Review (*NYT NWR*).

Oil and Gas J. (*OGJ*).

Petroleum Intelligence Weekly (*PIW*).

Petroleum Press Service (*PPS*).

Platt's Oilgram News Service (*PONS*).

Platt's Oilgram Price Service (*POPS*).

Wall Street Journal (*WSJ*).

LESSONS FROM AND FOR THE CURRENT ECONOMIC EXPANSION

Three Lessons

By PAUL W. MCCRACKEN*

A review of experience during the current expansion leaves an economist sympathetic with Matthew Arnold's metaphor about that "huge Mississippi of falsehood called history," and at the same time uneasily feeling that Santayana had a point when he observed that those "who cannot remember the past are condemned to repeat it." Experience during the current expansion has been a blend both of surprises and of economic relationships once again validated.

The latter should not be overlooked. Fiscal and monetary policies did cool the boiling inflation, with about the usual lags, after 1969. The price level did respond. The acceleration in the rate of inflation was halted, and then it began to recede. The turn-about in these policies in 1969 did generate a new expansion. And the over-heated expansion of 1973 is not out of character with the course of fiscal and monetary policies in 1972.

A certain back-to-the-drafting-board modesty on the part of our profession is, however, in order as we review the lessons to be drawn from this experience. Of the numerous candidates, my own comments will center around three. To some extent they are interrelated, but they are worth commenting upon seriatim.

The first is that we need to take a much more complex approach to estimating the upper boundaries of our capacity to produce, beyond which additional demand pressures would increasingly express themselves in inflation. While we all know that a point estimate of capacity implies a spurious precision and that the significance of the average unemployment rate varies with changes in the labor force and other factors, these overly simplistic guides continue to wield an undue influence on demand management policies. The current expansion has been a dramatic case illustration of how misleading these indicators can be. The gap between potential and actual output was estimated to be 3.2 percent of real *GNP* for the fourth quarter of 1972, and our productive potential was then estimated to be enlarging at the rate of 4.3 percent per year.[1] At the beginning of 1973, therefore, we presumably had room for a rise of about 7.5 percent in real output.

Obviously we had nothing like that kind of elbow room for expansion. We will have achieved somewhat over a 4 percent expansion in output during 1973 and then only with heavy pressures on the price level. And shortages of materials and plant capacity provide eloquent evidence that the economy has been pressed hard this year against the practicable upper limits of its capacity to produce.

If we look at the pressure problem more broadly we find that there were warning signals a year ago of pressures building in the economy. The proportion of companies reporting slower deliveries, for example, had risen during the expansion (which was to be expected), but the striking thing was that by October 1972 it had already reached levels seen in only three other years during the last two decades. And by early 1973 it had reached record levels. The proportion of companies reporting commitments of sixty days or longer for production materials was by year-end 1972 higher than at any time in fifteen years.

---

* Edmund Ezra Day University Professor of Business Administration, University of Michigan.

[1] *Business Conditions Digest*, 95.

A second lesson from and for this expansion is that external developments are obviously destined to have a much larger influence on the course of the domestic economy than has been true in the past. When we examine our own economy during 1973 we see the extent to which swings in our external payments have affected the domestic economy. And the recent expansion vividly illustrates this point. While historically exports have been 4–5 percent of our *GNP*, the rise in merchandise exports from the third quarter of 1972 to the third quarter of 1973 was equal to 17 percent of the rise in *GNP* and equal to an almost incredible 32 percent of the rise in goods output.

Imports, of course, rose also during that year, but the net impact of external transactions did exert a major influence on the domestic economy. The demand for domestic output in nominal terms during the first three quarters of 1973 rose at the rate of 10.5 percent per year, but the sharp change in our trade position added another 1.3 percentage points to the rate of growth in the total demand for domestic output. An increment of demand of that magnitude added to a rate of growth that already had the domestic economy hard-pressed was bound to produce a leveraged effect on the price level.

This all means two things. It obviously requires a more conscious and explicit effort to take external transactions into account in prognoses about the course of the domestic economy and in the management of domestic economic policy. More important, the direction of needed new machinery for international economic policy must be substantially altered. A larger proportion of our formal efforts directed toward new monetary arrangements and freer trade needs to be devoted to more explicit attempts to consult and coordinate domestic economic developments and policies among the more advanced economies—roughly, the Organization for Economic Cooperation and Development (*OECD*) members—or international disturbances will put into jeopardy not only prognostications about domestic economic developments but also appropriate policies for demand management.

My final comment has to do with programs having a direct concern for price and wage developments. This is a vexatious question. Our own history in recent years seems to fit the *OECD*'s characterization of international experience more generally. In its 1970 report, *Inflation: The Present Problem*, it observed that:

> The record seems to show fairly clearly that price-incomes policy of both the national bargain and guidelines variety has succeeded in restraining price rises for periods of up to one to two years. . . . To what extent any long-term success has been achieved, is, however, more doubtful and controversial. [p. 37]

We have had in the United States a stiff program for price-wage restraint, far more severe than is usually implied by the amorphous term incomes policy. Even Phase III had far more bite than the conventional guidelines or monitoring system that have frequently been commended to the managers of policy. Even so, these have not shown us the way to price stability.

Moreover, international experience seems to suggest that these programs also produce growing imbalances in the economy, at least to the extent that they have a bite, after a year or so. We now see these all around us. Artificially large exports of materials, induced by a disequilibrium between external and internal prices, produce artificial shortages at home. A low-profit expansion has resulted—with corporate profits as a percentage of national income far below levels of the mid-1960's and with real corporate profits in the aggregate not even regaining their 1964–66 average until late in 1972.

The immediate lesson to be drawn here is that our own wage-price program is now well beyond the point where the curve of its diminishing effectiveness has intersected the curve of rising distortions. The time has come for this program to be given an honorable discharge.

For the longer run this experience seriously raises another question. Would the economy perform better if, instead of accepting the economic distortions that inevitably cumulate from a price-wage program of doubtful long-run effectiveness, we allowed whatever shortfall from price stability we cannot prevent by using aggregative tools to express itself as open inflation? Experience during the current expansion is making an affirmative answer to that question look more persuasive.

REFERENCES

Inflation: The Present Problem, A Report by the Secretary General of the Organization for Economic Cooperation and Development, Paris 1970.

U.S. Department of Commerce, *Business Conditions Digest,* Nov. 1973.

A Strange Expansion

By William J. Fellner*

This has in many ways been a peculiar expansion. One way to describe its strange characteristics is to say that qualifications which in conventional presentations tend to be relegated to footnotes have recently acquired great importance, and some central propositions which in conventional presentations receive most of the emphasis have recently deserved not much more than a footnote. I will use three illustrations of what I have in mind. They all relate to experience which I believe should be kept in mind in the future though their importance may not be the same in successive cyclical expansions. They all are rooted in the fact that our rapidly expanding economy ran into shortages of specific types of resources with unexpected suddenness.

I. Potential Output and the Full-Employment Budget

It has been emphasized all along that the upward-sloping line representing so-called potential output may change its slope from time to time. In the past little attention was paid to the fact that a change of the slope may be associated with the fact that a given *general* unemployment rate may correspond to different degrees of labor market tightness at different times, but of late this complicating factor also has received attention. The fact that the slope of the so-called potential output path can change because of changing rates of increase of the labor force and of output per manhour has received attention all along.

What I think has usually not been taken into account—except perhaps as a somewhat pedantic refinement—is that the position and slope of the long-run output path are apt to depend on *how* the short-run movements of the economy are "engineered" by monetary and fiscal policy. I feel that the point has been driven home by our recent experience.

In 1972 and early in 1973 our short-run expansion rate was very steep. Overexpansion became all the more pronounced because we ran into specific supply limitations of an unusual and unexpected kind. Subsequently, the economy needed to be "cooled." Our growth has slowed and, for a while, will continue to slow to less than its long-run average rate even in the event that a recession in the conventional sense is avoided. In such circumstances, it may be advisable for monetary and fiscal policy to recognize the necessity of going through a phase in which the economy operates for a while at a higher unemployment rate than that developing subsequently. In our case it would in all probability have been necessary to do so even aside from our present energy shortage. To me it seems an exercise of little significance and even of doubtful logical validity to try to find out what the deviations from "potential output" will have been during the phases of development which I am describing. The alleged potential in question describes no norm for such a phase, e.g., for 1974. In such circumstances no good purpose is served by endowing a trend line with more meaning than it has by definition.

What we *can* properly say is that given the imperfect state of foresight, policy makers were trying to adopt policies which by their lights would place the economy on the best possible short-run path, and that they also had some conception of a much smoother and more desirable path on which it might be possible to move from some future time on if at that time foresight should prove more adequate. But it has become clear that the properties of any such hoped-for future growth path depend importantly on short-run movements in the relatively near future. At all times one finds oneself in a situation in which the imperfections of past foresight have led to various structural maladjustments which can be overcome only gradually, and failure to overcome these maladjustments would lead to a significant worsening of the properties of any feasible future path. That path does not live a life of its own but becomes established in part by how (and sometimes with what *detours*) we are trying to move toward reasonably stable long-run growth when we experience disturbances.

---

* Member, President's Council of Economic Advisers.

These considerations suggest the need to qualify the fiscal-policy principle expressing itself in full-employment budgeting. In general they have made me skeptical toward principles implying that policy makers can first determine the properties of a long-run potential output path and *then* set the fiscal decision variables in such a way as to have a prompt and automatic push develop in the direction of that path whenever we are not on it. The path in question may prove a harmful figment of the imagination if we pay too little attention to the timing of the moves intended to lead to a reasonably smooth long-run course and to the methods used for getting there.

On some occasions in the past it would have been unnecessarily pedantic to stress these weaknesses of a conventional analytical framework. Not so on the present occasion, and this should be remembered in the future.

II. Domestic Price Effects of Exchange Rate Adjustments

It has frequently been asserted that in a country with a small foreign trade sector, changes in exchange rates are apt to have small effects on the domestic price level. A qualification has not been overlooked but its potential importance has usually received little attention. This time the qualification has proved very important.

If significance attaches to that part of the domestically produced and domestically absorbed output which consists of close demand substitutes of exports or of imports, then no substantial price differentials will establish themselves between the internationally traded goods and components of the domestic output to which we attribute importance. As for the United States, a large fraction of our exports and of our imports does have close substitutes on the demand side among the domestically produced and absorbed goods. It is, of course, true also that very significant sectors of our economy are producing goods and services the dollar prices of which will behave differently than those of the traded goods when exchange rates move. However, the prices of domestic *agricultural products* as well as of domestic materials used in generating *energy* do move together with traded goods. In a period in which the exchange rates of the dollar had become significantly reduced, this linkage with internationally traded goods added a good deal to the upward price pressure on commodities that were getting scarce anyway.

In spite of their moderate weight in our various price indices, the prices of farm products and of sources of energy became the main culprits in the recent inflationary scene. They acquired great importance for the interpretation of recent developments because their prices have risen exceedingly steeply. Hence the price link between these specific categories of domestically produced and absorbed goods on the one hand and internationally traded goods on the other also became very important, despite the small weight of foreign trade in American economic activity.

III. Real Wages and Unit Labor Costs

Economists have all along been aware of the fact that results obtained from two-input models—"labor and capital" models—are subject to qualifications, though in the present century we have become less reluctant to employ such models than, say, Ricardo would have been. Furthermore, economists have all along been aware of the distinction between real wage rates and product wage rates, though normally we would probably have concluded—at least in an acceptable first approximation—that if in the private nonfarm sector real compensation per manhour increases no more than productivity, then employers are compensated for the money-wage increase by the sum of the productivity increase and the price increase.

It would be no surprise to anyone in our profession that such a calculation may not come out precisely right, since it involves a crude approximation. But I think the "approximation" has rarely proved as misleading as recently. During 1973 real compensation per manhour will probably have declined somewhat in the private nonfarm sector, while productivity will have increased. At the same time money compensation per manhour in that sector will have increased steeply from the employer's point of view. It will have increased in a higher proportion than the sum of the sector's productivity increase and of *its* price increase. This may well turn out to have been the case even if the relevant productivity increase were interpreted as some reasonably defined average of the private nonfarm sector's *long-run* increase in manhour output and of the simultaneous productivity increase. At any rate not even temporarily will the cyclical increase in the profit share have restored it to its normal cyclical high.

The explanation is provided by the fact that real compensation per manhour would have risen steeply during the year if in computing real compensation we corrected the increase of money compensation not by the total consumer price index but merely by the nonfood component of that index (or, what is the same thing, if the food component had risen in the same proportion as nonfood). Even real wage rates in the narrower sense, excluding fringes, would have risen considerably when so calculated. Nonfood consumer prices and the private nonfarm deflator have risen very much less than consumer prices as a whole or than the *GNP* deflator as a whole. Of course, it makes no sense to exclude food prices when looking at the problem from the worker's point of view; but from the typical employer's point of view in the American economy, a steep increase in the prices of farm products and of finished foodstuffs is not an offset to rising money-wage costs. In the long run any appreciable discrepancy is very unlikely to persist between the rate of increase in money compensation per manhour in the private nonfarm sector and the sum of the sector's productivity increase and the price increase of *its* products.

However, there is reason to expect that the significant difference between the interpretation of rates of change when seen from the typical worker's point of view and when seen from the nonfarm employer's point of view will not last very long. This difference too belongs among the peculiarities of a recent period, though it should serve as a reminder that we do not live in a two-factor world.

Indeed, while the relative prices of scarce sources of energy will be rising, this will create a situation which is similar "in essence" to that just discussed. "In form" the two phenomena are not similar because the fact that the change in labor compensation has been disappointing from the worker's point of view and yet steep from the typical employer's point of view has resulted mainly from a significant rise in *farm* prices, while the prices of crude oil, coal, etc. are included in *nonfarm* prices along with the prices of industrial products. Yet "in essence" the two phenomena will show similarity because a substantial rise in the prices of scarce energy sources affects the cost of living without resulting in a corresponding increase in the price which is implicit in the value added in industrial activities. In this regard the role of the crude oil price or of that of coal is very similar to that of farm prices.

Lessons from the Current Economic Expansion

By PAUL A. SAMUELSON*

I shall take literally my assignment on this panel and discuss what lessons we have learned from recent macroeconomic experience. This means noting some surprising departures from previous patterns of experience. But it also means pinching ourselves and asking, What went right?

Here, then, in ten minutes are ten lessons learned or to be learned, which is at the rate of one lesson a minute.

Lesson 1. *Economists cannot forecast the future with precision.*

Lesson 2. *Neither can anyone else,* and the forecasts made by professional economists are systematically better than any made by brokers, bankers or businessmen. (Together, these lessons might constitute the "random dart" theory of getting a forecast: throw a coin at any one of the couple of dozen leading practitioners of the forecasting art, and you'll do about as well as if you try to pick out the very best one. In fact, it won't matter much where your coin lands, you'll get almost the same forecast anyway—which leads to the next lesson.)

Lesson 3. *Economists can succeed in forecasting like each other.* Roy Blough's dogma still holds. Economists are like eight Esquimaux in one bed—the only thing you can be sure about is that they are all going to turn over together. I may add as a corollary to Blough's Law: most economists are going to be all correct together, or all wrong together. The dispersion between forecasts by a score of university, bank, government, and corporation analysts is much less than the dispersion of their best guesses around where the future reality will lie.

Lesson 4. *Some economists manage to forecast worse than others.* But a Darwinian process of selection—often self-selection when one deals with rational men with alternative uses for their leisure time—does serve to eliminate those whose comparative advantage lies in cosmology or in futurism. After a man has said a dozen times, "If this forecast doesn't turn out correct, I'll hang up my gloves," he finally does. However, there are exceptions to this rule: some masochists make a nice living out of being wrong in an interesting way, predicting every other year that the Dow-Jones average will go down to 200, where it will meet the dollar price of gold. Businessmen, who like to have their adrenalin run listening to a ghost story, will pay as much to listen to these ever-losers as to informed analysts. But my Darwinian proposition becomes irrefutable when I base it on the tautology that the born losers who stay in the forecasting game cease to be regarded as economists by members of our guild.

Lesson 5. *An automatic computer can forecast better than an official government agency.* But an analyst with judgment can do better than an automatic computer. This is borne out, not only by the rather sad performance of the reduced-form monetarist model of the Federal Reserve Bank of St. Louis, but also by Ray Fair's audit of his own Princeton model, which does without any inputs of "judgment" and which thereby runs up a larger mean square error than the leading judgmental-computer models of the Wharton, Michigan, *DRI*, Chase, or other type.

---

* Massachusetts Institute of Technology.

Lesson 6. *Judgmental analysts cannot do without the computer.* So much of the information that comes to any man of judgment these days has been massaged by a computer that the era is past when one could have a fair contest between a nonanalytical observer such as Sumner Slichter and a giant post-Keynesian macro computer model. Were Slichter alive today, he would make his forward strides of shrewdness while riding on the escalator of the computer.

Lesson 7. *Economists can forecast (well, almost forecast) everything but prices.* In this sense they are the reverse of Oscar Wilde's cynics, who know the price of everything and the value of nothing. When it comes to next March's price of wheat, why should any Ph.D. be able to forecast that better than people who know a lot about the crops and milling industry and who have a lot of money riding on the outcome of their best guesses? All the things that are easy to forecast—that is, easy enough for a mere professor to foresee—can be expected to have already been taken into account by the speculative market. Indeed, if there were in Las Vegas or New York a continuous casino on the money *GNP* of 1974's fourth quarter, it would be absurd to think that the best economic forecasters could improve upon the guess posted there. Whatever knowledge and analytical skill they possess would already have been fed into the bidding. It is a manifest contradiction to think that most economists can be expected to do better than their own best performance. But I am saying something more devastating than that. I am saying that the best forecasters have been poor in predicting the general price level's movements and level even a year ahead. By Valentine's Day 1973 the best forecasters were beginning to talk of the growth recession that we now know did set in at the end of the first quarter. Aside from their end-of-1972 forecasts, the fashionable crowd has little to blame itself for when it comes to their 1973 real *GNP* projections. But, of course, they did not foresee the upward surge of food and decontrolled industrial prices. This has been a recurring pattern: surprise during the event at the virulence of the inflation, wisdom after the event in demonstrating that it did, after all, fit in with past patterns of experience. I know of no exception to this generalization. Monetarists have generally done no better in this regard, although some have thought that there is somehow a closer link between M and P than between M and money $GNP=P\times Q$. To be sure, some have been less unlucky in their guesses than others, but we all know what the corollary of that is. What seems to have thrown economists off is the fact that the Phillips Curve seems to have worsened, both in the eyes of those who believe in the Phillips Curve and in the eyes of those who don't (and use a different language to describe the same disappointing effect upon wages and costs generally of each degree of stagnation). Improvement in the rate of inflation is a mirage that economists see eighteen months ahead; but just as the railroad tracks never seem to meet as they appeared to do ten miles back, so that return to "normalcy" never seems to occur—not even after we have euphemistically redefined normalcy to permit of 2.5 percent annual inflation or more. Paradoxically, as soon as the analysts stopped underestimating wage increases, they began to underestimate general price increases. One suspects that two things are involved: wishful thinking and perhaps a change in the structural difficulty of the economy in having steady or slowly growing price levels. When their predictions go wrong, people usually find alibis in contrafactual predictions incapable of being falsified or corroborated. Thus, some say, "It is price controls that cause the error in my forecasts (in either direction!); but, of course, we shall never know what the world would have been like without those price controls." Or they say, "If the Fed would do something (which the odds are against its doing), then the reality would agree with my forecast." Much of the disagreement for policy, but not all, comes from differences between advisers on how much sacrifice of short-term welfare they are willing to invest in now and in the near future for some hypothetical better behavior of the economy in the more distant future. Those with a low rate of time discount, who will cheerfully recommend a flyer in sadomasochistic austerity, are likely by the Darwinian process of politics to find themselves not in power for the period of time they say their therapy requires.

Lesson 8. *In recent years the U.S. economy has been remarkably unprone to swings of inventory accumulation such as characterized our past history.* Those who prior to the energy crisis predicted a genuine recession for 1974 were counting on such an inventory slingshot effect to give them their downturn. But they may have to look elsewhere for their recession.

Lesson 9. *A microeconomic event such as an Arab oil boycott can loom large as a macroeconomic depressant.* Most experts now think that if the boycott is more than a charade and a passing thing, real output will decline in the first half of 1974. But we now realize that our Fisher-Keynes macro models do not tell us how to handle such a microeconomic restriction on supply and productivity. It is not even clear that we should call such a downturn a recession, since it may not involve much wastage between our actual *GNP* and our producible *GNP*. However, to the extent that gasoline scares hurt the auto business and the suburban building market, some of the same secondary effects of conventional recession will become operative, calling for some of the conventional lean-against-the-wind policy measures.

Lesson 10. *The conventional wisdom that currency depreciation will improve the balance of payments seems in 1973, at long last, to receive some support from the factual evidence.* There have been more important things to worry about in recent years than which way the dollar would move in a regime of somewhat-flexible exchange rates.

I've still not drawn all the lessons taught by our recent times. Perhaps I may be permitted to give one final lesson, to grow wise on, so to speak.

Lesson 11. *Events of the last half dozen years have shown us how much economics remains an art rather than a science.* Economics is exciting because what we study is hard, not easy. Once again, I think, experience has taught us the hard way that eclecticism in economics is, to paraphrase Justice Holmes, not so much a desirability as a necessity.

A Plethora of Lessons from the Recent Expansion

By OTTO ECKSTEIN*

Judged by conventional business cycle standards, the recent expansion was so strange that one has to be careful not to impute every newly uncovered problem to the indefinite future. Things will not always be this bad, although one must admit that the incredible projections of the Club of Rome model are unrolling with a speed which seems to have us halfway to Doomsday just two years after the model's publication.

The pattern of the expansion. The economy began to recover from the saucer-like recession at the beginning of 1971. Initially, real growth was slow because business remained extremely cautious in its inventory and plant and equipment policies, consumers continued to be frightened by inflation, and our international trade position continued to deteriorate. But by midyear, and particularly after the President's New Economic Program was announced, the economy took off in a growth spurt which produced an average rate of real growth of 7.2 percent for seven quarters. The expansion was led by housing, which benefitted from the end of the 1969–70 money crunch. The consumer sector began to show large gains in the second half of the year. Business spending attitudes improved at the beginning of 1972 and brightened once more in the closing months of that year. During 1973 all private sectors except housing continued to advance and combined to produce another year of high growth.

The growth of the seven quarters of strong expansion was clearly unsustainable. In terms of manpower, the gap of unutilized resources was not closed, but by early 1973 manufacturing capacity proved to be the bottleneck to further expansion. The agricultural price explosion, which also infected the industrial commodity exchanges, made the inflation acute; real purchasing barely rose; consumer debt was grossly excessive. The stock of housing and commercial buildings had become too large. The downturn was inevitable. As the contractionary phase of the business cycle began in the second quarter, the Federal Reserve belatedly tightened up on monetary policy, created a brief money crunch, and thereby worsened the decline. As the finishing touch to this ill-starred cycle, the energy crisis broke upon us.

Lesson 1: *Less fine-tuning for demand management.* Why did this business cycle develop into a classic, unsustainable boom? The causes lie in the previous cycle history and in the economic policies. The money crunches of 1966 and 1970 depressed the housing stock, producing an exceptionally buoyant prospect for the housing industry. Further, the massive tax reductions of 1969–70 boosted consumer and business purchasing power, but the uncertainties of the Vietnam War and the inflation slowed the response of spending to tax reduction even beyond the normal time lags.

In retrospect, it is also clear that fiscal and monetary policies were too stimulative during the seven quarters of the rapid advance. The full-employment budget, even if calculated on the obsolete 4 percent unemployment basis, was put into substantial deficit by the tax reductions. The average full-employment budget during the seven quarters of the boom was in deficit by $1 billion compared to an average surplus of $5 billion in the entire period since 1955. This error in fiscal policy can be attributed to two factors: first, it was not appreciated how long it would take before the tax reductions of 1969 and 1970 worked their way into the spending stream and created their multiplier effects. Thus, the New Economic Policy of August 1971 was added on top of fiscal moves that were just beginning to take hold. Second, the administration's political need for a quick return to prosperity in the year before election led to the resolution of policy issues on the side of stimulus rather than restraint or neutrality.

---

* Data Resources, Inc., and Harvard University.

Some of the same forces were at work on the side of monetary policy. As Arthur F. Burns took over the leadership of the Federal Reserve System, policy switched to the management of monetary aggregates. In the initial years of the new approach, there was an inevitable learning process, and some mistakes were made. In 1971 and early 1972, the aggregate approach produced short-term interest rates that were clearly inappropriate for an inflationary economy. These rates were produced by international short-term money flows and other technical factors, and probably should have been contravened by monetary policy.

More fundamentally, history gave little guidance to the selection of the appropriate target rate of increase for the money supply. In 1971 the money supply rose by 6.6 percent, in 1972 by 8.3 percent, measured December to December. Rates of increase of 6 to 7 percent were commonly considered to be the normal rates of increase given the inevitability of some substantial inflationary process. Thus the increase of 8.3 percent during 1972 was considered excessive, but the margin of error was believed to be small.

In hindsight, it is clear that this monetary policy was too expansionary. Personal saving was very high during much of this boom, and corporate cash flow advanced at record rates, an average annual rate of 17 percent during the three years, 1971–1973. Thus, the supply of private saving was enormous and, when combined with a relatively easy monetary policy, produced an overabundant supply of long-term investible capital. As a result, financial institutions overencouraged real estate developers and other entrepreneurs to organize projects, and a boom developed in the construction of office buildings, shopping centers, apartment houses, and other kinds of construction. Even in the single family home market, the overabundant supply of capital combined with rising real family incomes to produce a level of construction far beyond the trend requirements.

Lesson 2: *More planning for physical capacity.* The shortages of processed industrial materials have a more explicit economic foundation. Virtually every major processed material, such as metals, chemicals, paper, lumber, rubber, glass, was in short supply toward the end of the growth spurt. When devaluation and domestic price controls led to an export boom, the worst shortage situation ever seen in peacetime was created.

This lack of capacity is the culmination of a fifteen-year historical process. Many of these industries overexpanded in the boom of 1955–57. The overvalued dollar of the 1960's did serious damage to their competitive position abroad and gave foreign producers an easy entry here. Since the mid-1960's, these industries have been out of favor with the investment community; their rates of return have been relatively low; the price controls threatened to keep them low; and the difficulties of adjusting to new and uncertain environmental standards further delayed investments. As a result, the actual growth of capacity of these industries was low, and the economy ran out of primary processing capacity long before resources as a whole were fully utilized. The shortages of 1973 were worse than in 1966: yet unemployment at that time was 3.75 percent, compared to 5 percent this time. Thus, physical and human resources have become disproportionate. This bottleneck, in turn, creates shortages in the succeeding stages of production: while the machinery, computer and other industries have the ability to produce increasing output in their own factories, they are unable to purchase critical inputs. The shortage of finished products, felt all the way to the retail stage, produces inflationary pressures.

Neither business nor government was alert to these problems. Nor did the economists show more foresight. We are all accustomed to an economy dominated by demand forces and take the solution of the production problems for granted.

With demand driving up prices and the rates of return, these industries will expand their capacity. In some industries, the shortages can be ended in a year or two. In others, it will take much longer. Economic analysts and economic policy will have to give weight to these problems.

Lesson 3: *If you have controls, make them believable.* When the President abandoned his policy of reliance on market adjustments in the summer of 1971 and instituted the wage-price freeze of Phase I, he began the first of a series of sharp swings in controls policy. The succeeding phases represented a continuing tension within the administration between advocates of direct controls and their ideological opponents. While disagreement can be an aid to policy in the formation stage, the economy was harmed by the actual oscillation in policy as the program staggered through the various phases I to IV. The swings of policy were too sharp, too frequent, too unpredictable, creating an economic climate which was unnecessarily insecure, and finally producing disenchantment by the public. The President's own lack of involvement in the program also cut its credibility. He set the basic policies, but provided little moral leadership to identify public compliance with the national welfare, nor did he give much practical backing to the operation of the programs. As a result of this history, the controls probably cannot be used effectively again as an anti-inflation weapon for quite a few years

Lesson 4: *International interdependence has become a major source of instability.* The rapid growth of trade has made the separate national economies, both advanced and less developed, far more interdependent than they were before. Even for the United States, the ratio of imports to *GNP* has risen from 4.6 percent in 1953 to 7.4 percent in 1973. Our increasing dependence on imported materials and energy sources makes trade even more critical to our prosperity. The free movement of currencies and capital has created other new links of interdependence. As a result, even domestic business cycle analysis has to consider the rest of the world in detail.

This interdependence throws up a host of new problems for policy. First, trade policy no longer can focus mainly on import barriers, but has to deal with the sudden and unpredictable imposition of export controls. The rules of the General Agreement on Tariffs and Trade have little to say about this matter, and there is no international forum or institution in which nations are brought together to develop a reasonable set of working rules. During the last period of commodity shortages, in the Korean War, institutions were created for this purpose. Second, a system of fixed exchange has clearly been made impossible. Floating rates are considered temporary and are operating without principled rules for government intervention in the exchange markets. Yet the risks of competitive appreciations to fight inflation or competitive depreciations to fight unemployment are large. Finally, foreign exchange flows and holdings created by the energy crisis will test the existing payments mechanisms.

Lesson 5: *Policy machinery is obsolete.* The basic machinery for formulating economic policy was designed for a situation where demand management was the main problem. The Council of Economic Advisers, the Troika, and the Quadriad were designed to advise the President on short-run macro policy. The Office of Management and Budget is good at policing spending policy. The Treasury can analyze and advise on tax and international issues. But the actual economic problems of the last expansion went far beyond these areas. If commodity shortages, capacity constraints, and alternative energy sources are among the central tasks of economic policy for the intermediate term, the economic policy machinery must reflect this state of affairs. Serious long-term planning issues are posed. The task of policy also becomes far more operational. The decisions are smaller but more frequent and detailed. At this time, the Cost of Living Council is concerned with the micro aspects of price and wage stability, and it is also concerned with domestic adequacy of supplies. The Federal Energy office seeks to be all-powerful in dealing with its particular physical, price, and international aspects. Other international commodity problems are a loose ball on the field. On a *de facto* basis, Treasury Secretary George Shultz provides the policy coordination in his many roles, but no one man can be the substitute for a rational policy process on topics that involve so much detail.

Perhaps the biggest challenge to economic policy over the next five years is to develop the principles, the analytical techniques, the information base, the policy formation machinery, and the instruments that will be needed to solve these problems. Since our reliance must continue to be primarily on the productive activities of private enterprises, and since the United States has no healthy tradition of business-government collaboration in peacetime, the design of such machinery will be very difficult.

TWO YEARS OF WAGE-PRICE CONTROLS

The Legacy of Phase II Price Controls

By ROBERT F. LANZILLOTTI AND BLAINE ROBERTS*

President Nixon's decision in July 1971 to invoke the Economic Stabilization Act is generally regarded as a surprising and drastic reversal of the economic game plan for 1971. Taken in peacetime by a Republican President ideologically opposed to controls, it was, of course, a surprise. Viewed from the perspective of economic policy development in the United States, however, the Freeze and Phase II represent more an extension of the guideposts of the 1960's than a sharp break with earlier economic policy. Whether or not the controls decision was economically sensible will be argued indefinitely. The fact remains that from the standpoint of political imperatives, the 1971 game plan was no longer tenable.

The main purpose of this paper is to reflect upon the Phase II experience and to answer from an inside perspective some of the questions that have arisen. (1) Was Phase II an economic "sham," a controls placebo for the public? (2) What were the economic effects of Phase II? (3) What are the basic lessons and legacies of the Phase II experience?

I. The Effectiveness of Phase II

It is rash to presume that the effectiveness of Phase II could be assessed definitively so soon after its abrupt termination in January 1973. Nevertheless, we shall attempt to identify and briefly assess the economic issues generated by the experiment, as we view them, conceding that a rigorous overall evaluation encompassing more "phases" and "stages" and taking a longer view will no doubt look different.

Several econometric models used to appraise Phase II, including our own research, indicate that the effects of the six quarters of controls include: (1) an average reduction in the annual rate of inflation (at the aggregate, final consumption level) of about 2.0 percentage points and (2) a small (or ambiguous) alteration in the rate of change in wages.[1] Assuming these estimates are accurate, one would expect a concomitant effect upon other macroeconomic variables. More specifically, assuming (a) nominal *GNP* was fixed by monetary and fiscal policy prior to the Phase II experiment,[2] and (b) that rates of change in output, employment, and productivity basically follow Okun's law, we estimate (3) an increase in real *GNP* of some \$15 to \$22 billion (associated with the 2.0 percentage point reduction in the rate of inflation), (4) an average reduction in the rate of unemployment of 0.2

---

* Professor and assistant professor of economics, respectively, University of Florida.

[1] Econometric studies showing approximately these results include those by Robert J. Gordon; Bradley Askin and John Kraft; Robert F. Lanzillotti and Blaine Roberts; John Kraft and Blaine Roberts; and Charles Guy, John Kraft, and Blaine Roberts. Descriptive analyses consistent with the above are those by Barry Bosworth and Joel Popkin. Edgar Feige and Dennis Pearce, using an *ARIMA* model, as distinct from a structural equation model, argue that there was little effect on inflation of both wages and prices.

[2] Equivalently, the assumption is that aggregate demand has a unitary price elasticity and was unaffected by controls.

percentage points (or an increase of about 160,000 jobs), (5) an increase in the average rate of productivity of about 1.0 percentage point, and (6) a zero or ambiguous effect on the rate of change in corporate profits.[3]

On the cost side, we estimate that *direct* federal expenditures and costs incurred by firms to comply with the regulations amounted to about $0.8 billion annually. While there were some problem areas arising from the booming world demand (e.g., hides and lumber), there were no serious inefficiencies created by Phase II controls, and we believe a generous estimate of the cost of distortions caused by controls to be $0.5 billion. Thus, the economic "benefit-cost" ratio from Phase II is large (between 11.7 and 17.5).

Phase II regulations apparently did not affect either the functional or size distribution of income. Changes in functional shares that did occur are easily explained by the stage of the business cycle. Furthermore, price indices constructed for high, medium, low, and retired incomes show equal rates of inflation.[4]

In terms of the specific effects on rates of inflation by industry, however, significant differences were found. Phase II evidently exerted little or no effect upon rates of change in prices in the manufacturing sector;[5] most of the reduction in the aggregate rate of inflation appears to have been concentrated in the service and retail-wholesale trade sectors.

The econometric results cited are open to a number of criticisms and qualifications; the temptation to attribute the observed results entirely or largely to Price Commission (*PC*)-Pay Board policies should be resisted. Likewise, it is highly debatable whether the same degree of effectiveness disclosed by the studies could have been continued in 1973. Nonetheless, even though the program could not be regarded as the unqualified success claimed by some, it certainly did not result in the unmitigated economic fiasco of disruptions and shortages predicted by others.

II. The Cruxes of Phase II

The results reported above provide a clinical view of the program's effectiveness, but other questions of interest to economists deserve comment. We believe some better insight into the nature of these matters can be provided by viewing the program in terms of two of its major issues or "cruxes" that were fundamental to the success of Phase II: (1) the generation and maintenance of a "stabilization ethic" and (2) the fundamental administrative issue we describe as the "control of controls." These are not to be regarded as the sum total of the basic issues or the principal insights gained, but constitute a useful vantage point from which to appraise the experience.

A. *Creating a "Stabilization Ethic"*

It has been observed that Phase II was "successful" because it had everything going for it—an economy recovering from a recession, the presence of underutilized capacity, the prospect of productivity gains, and the competitive market forces characteristic of this stage of the business cycle. Thus, the economic climate was propitious, but equally significant was the sympathetic receptivity of business, labor, and the public generally to the development of the anti-inflation or "stabilization ethic." This may well have been a short-run, one-shot attitude. Nonetheless, during most of Phase II the atmosphere generally reflected support for and compliance with regulations designed to reach the announced stabilization goals. The effects at the aggregate level are reflected in the econometric studies cited, even though particular results cannot be traced to any specific regulations.

---

[3] Full details of the econometric models are presented in Kraft and Roberts.

[4] Details are in Lanzillotti, Roberts, and Hamilton.

[5] Details are contained in Lanzillotti, Roberts, and Hamilton, and the specific econometric models are presented in Guy, Kraft, and Roberts.

In January 1973, however, the public was informed by the President that Phase II was not good for the economy after all because controls tend to become narcotic. The two regulatory bodies that had come to represent the touchstones of price and wage stability were terminated, and Phase III was launched with as much fanfare as one might find with the opening of a mortuary. More than that, Phase III, as a controls program, had no high-level sponsor, or at best it was a tenuous and ambiguous sponsorship. Stated differently, to the extent that Phase II officials successfully generated a stabilization "ethic," it proved entirely too fragile to withstand the abrupt and confusing "dash back to market" represented by Phase III.

Whether the effectiveness of Phase II controls could have been maintained into 1973 is, of course, debatable, since it depends upon answers to still other contentious issues: (1) What are the fundamental causes and processes of inflation? (2) How do controls alter this process both at the microeconomic and macroeconomic level? (3) How do the current economic conditions affect the answers to (1) and (2)? While economic theory and empirical research are working on more accurate answers to these questions, we can make a few subjective observations with respect to Phase II's apparatus and policies.

In the latter half of 1972 general cooperation with the control program appeared to deteriorate, at least in particular sectors of the economy. Economic responses to the problems were repeatedly delayed, and the Cost of Living Council (*CLC*) blocked several important changes proposed by the Price Commission, creating an impression that there was a conscious and planned reduction of support for controls. Thus, the political atmosphere preceding the presidential campaign, characterized by reluctance to make needed changes, added to the growing sectoral dissatisfaction with the regulations.

Furthermore, if the stabilization ethic hypothesis correctly explains part of the effectiveness of Phase II, the way Phase III was announced and promulgated served to undermine this ethic. Specific stabilization goals were scrapped, directions and new regulations were unnecessarily vague, and the stress on price stabilization soon gave way to a race to raise prices. The annual rate of increase in the wholesale price index (*WPI*) of 22.4 percent and 8.0 percent in the consumer price index (*CPI*) between December 1972, and June 1973, when Phase III gave way to a price freeze, reflects—in significant measure, we think—the consequences of the abrupt policy shift in January 1973.

B. *The Control of Controls*

The "control of controls" involves questions about both external and internal administrative arrangements, namely: (1) the degree of autonomy afforded to the controls agencies within the organizational structure of the Economic Stabilization Program and (2) the internal administrative organization of the agency. The degree of autonomy of the Pay Board was raised and resolved in its favor by prospective labor appointees to that Board before the Phase II units were organized. On the other hand, although not raised initially, the degree of autonomy of the Price Commission became an issue during Phase II but was not fully resolved. There were several instances when Price Commission autonomy clashed with Cost of Living Council authority. This was due in part to the feeling of some administration officials that the Price Commission tended to overstep the bounds intended. Four instances of *PC-CLC* differences over policy matters indicate the nature of the problem.

(1) *The content of the initial regulations.* A provision effectively eliminating all prenotification requirements between November 15, 1971, and January 1, 1972, had been inserted into the draft regulations (evidently at the behest of a *CLC* member) without prior notice to the Commission. At the insistence of one Commissioner, this was revised to allow the Commission seventy-two hours to challenge proposed price increases. Without this reversal of *CLC* action, the price regulations would have been emasculated before they were given a chance to operate.

(2) *The PC's "declaration of independence" in refusing automatically to pass through any wage increase approved by the Pay Board.* The issue arose during the United Mineworkers' negotiations in late 1971, when despite pleas from the Secretary of Labor for advance agreement on a pass-through of labor cost increases, the Commission decided that it would limit justification for price increases from labor costs to the basic Pay Board standard of 5.5 percent plus 0.7 for fringes.

(3) *The food crisis of June 1972.* After the Commission had drafted a memorandum stating that the food price situation had to be resolved by means other than price controls, the *CLC* effectively neutralized the *PC* as a policy-making body by blocking any proposed substantive changes in price regulations.

(4) After the automobile companies had complied with *PC* regulations regarding proposed fall 1972 price increases and had reduced their requests below what the regulations allowed, company chief executives were summoned to the White House, where they were asked by *CLC* Director Donald Rumsfeld and Treasury Secretary George Shultz to reduce or withdraw their price increase requests.[6]

We do not cite these instances to censure *CLC* for doing what it perceived to be its job, but rather to ask whether a control agency (or agencies) can be autonomous, that is, largely free from political influence on its policies and actions. Put differently, given a decision to institute wage-price controls or a review board of some type, the question is which is to be preferred from the standpoint of public policy: (a) a quasi-autonomous unit or units like the Pay Board and Price Commission of Phase II; (b) a Cost of Living Council, as in Phase III and IV, directly responsible to the President or his deputy; or (c) some third alternative.

Basically, the answer depends upon whether a more highly coordinated system of economic policy, with powers vested solely and directly in the executive branch, is preferable to an independent (or quasi-independent) agency functioning under a system of checks and balances. If the United States has indeed turned the corner toward institutionalized direct control of prices and wages, the effectiveness of any particular form of controls will depend on how the issue of "control of controls" is resolved.

In this connection there were several instances in Phase II where the policies of the Commissioners were not fully implemented, or were even reversed by the Commission staff. This resulted from a combination of circumstances: the part-time status of Commissioners, the desire to resolve problems as quickly as possible, and the almost inevitable tendency of an agency's staff to exert control over policy —even if it means elevating procedures into policy. Such administrative details may sound trivial but as a practical matter are crucial to the development of effective policy. Full-time status for Commissioners would help minimize the problem, but some might feel it would serve only to provide more opportunity for Commissioners to engage in economic mischief.

---

[6] This was viewed as "arbitrary and discriminatory" by Lynn Townsend, Chrysler Corporation, in a post-meeting press release.

III. Legacies and Lessons of Phase II

European evidence indicates that controls soon become unmanageable and are discarded.[7] The demise of Phase II can be viewed in a similar vein. Nonetheless, the measured results of Phase II, the logic that controls in such a vast economy as the United States "appear" to be necessary in some sectors, and the general evolution of *U.S.* economic policy lead us to the following general conclusion: public choice as determined by sociopolitical processes, rather than purely economic considerations, suggests the need to institutionalize some form of permanent prior review of price and wage increases. Controls very likely will be taken off within the next few months and reimposed within the next few years in a much different form than those of Phase II or subsequent Phases. Our expectation is that "controls" of some sort will survive, in some unit whose primary function would be a "watchdog" agency over price stability. The most important role of such a watchdog would be its ability to influence public policy through interaction with other agencies of the federal government. Such an agency would not necessarily ask that prices be held down regardless of the prevailing economic conditions. For example, one of the lessons of Phase II is that currently it is perhaps best to completely decontrol the prices of petroleum.

A. *Administrative Structure*

In this regard the conditions leading up to the imposition of controls suggest consideration should be given to amending the Full Employment Act of 1946 to make "price stabilization" an additional national goal, and to creating a Stabilization Board with a fixed number (e.g., seven public members) appointed by the President, with Senate confirmation, for terms comparable to those of the Federal Reserve System or the Federal Trade Commission. Their function would be to monitor price and wage movements in key sectors of the economy with the right to limit, after holding public hearings, prospective *or actual* price and wage increases on the basis of predetermined norms or guidelines. Such an agency would be expected to assess and report periodically to the Congress the price ramifications of various governmental policies and private actions, thus including in the public debate the inflationary consequences of past or contemplated actions.

Details cannot be fully explored here, but two basic advantages of such a proposal absent in the current mix of governmental agencies are (a) political independence enabling the Board to elicit national support and (b) the creation of an agency whose *primary* public charge is price stability. Two fundamental disadvantages are (c) a loss of control by the executive branch and a lessened ability to coordinate economic polity by a single individual or office and (d) the creation of another bureaucracy with all the undesirable consequences (e.g., more red tape, federal intervention in private decisions, and the tendency of "control" agencies to become the pawns of the industries they are intended to control). While the disadvantages should not be minimized, the Phase II experience suggests such an arrangement can serve as a useful adjunct in developing effective economic policy.

B. *Operational Policies and Standards*

On a more specific level, certain basic lessons and suggestions for future controls emerge. In particular, if the principal contribution of controls is through "voluntary" cooperation[8]—i.e., the stabilization ethic—then:

---

[7] See L. Ulman and R. Flanigan.

[8] Voluntary cooperation is to be interpreted as analogous to the voluntary cooperation associated with

(1) Realistic, specific, announced national goals for wages and prices, coupled with appropriate exceptions and/or exemptions, are essential to supply the information necessary to create public confidence in the effort.

(2) Regulations must be flexible enough to adjust to different economic conditions over time and between sectors; but should not be arbitrary or be used arbitrarily.

(3) Political processes should not be used extramurally as a means of invoking special *ad hoc* standards. Put differently, the more apolitical the controls, the better such controls will be able to elicit national support.[9]

(4) Administration of the control program, relations with firms and unions, and public relations are all critical to the success of a program. Consistent policy closely controlled by full-time decision makers, as opposed to part-time controllers, is important to the development of sensible policies and rulings.

IV. Conclusions

Phase II and all the diverse assessments it has generated, including this one, will not settle the fundamental issue concerning the appropriateness and economic efficacy of direct controls, even on a short-run basis. The difficulties, complexities, and side effects no doubt reinforce arguments of opponents of controls that the economy is not a "tinker toy" to be plugged and unplugged into weird shapes. At the same time, the recurring inability of traditional economic policies to maintain durable price stability—for whatever reasons—also reminds us that, much as we might wish, the economy simply is not a smooth, uncomplicated machine without moving parts or friction. Somewhere in between must lie a path to rational and effective economic policy. We believe the decisions of August 15, 1971, mark a significant turning point in the history of *U.S.* economic policy in the direction of that much sought-after path.

REFERENCES

B. Askin and J. Kraft, *Econometric Wage and Price Models: Assessing the Impact of the Economic Stabilization Program,* Boston 1974.

B. Bosworth, "Phase II: The U.S. Experiment with an Income Policy," *Brookings Papers on Econ. Act.,* 1972, *2,* 343–83.

E. Feige and D. Pearce, "The Wage-Price Control Experiment—Did It Work?" *Challenge,* July/Aug., *16,* 40–4.

R. J. Gordon, "Wage-Price Controls and the Shifting Phillips Curve," *Brookings Papers on Econ. Act.,* 1972, *2,* 385–421.

C. Guy, J. Kraft, and B. Roberts, "The Price Control Experiment: Short Run Success and Long Run Failure? An Econometric Examination of Wages, Prices, and Profit Margins in Sixteen Manufacturing Industries," 1973, Univ. of Florida, xerox.

R. F. Lanzillotti and B. Roberts, "An Assessment of the U.S. Experiment with an Income Policy," paper presented at the Tulane Conference on Incomes Policies, April 1973.

——— **and M. Hamilton,** *Phase II in Review: The Lessons and Legacy of Price Control,* Washington 1974.

J. Popkin, "Prices in 1972: An Analysis of Changes During Phase 2," *Monthly Labor Rev.* February 1973, *96,* 16–23.

L. Ulman and R. Flanigan, *Wage Restraint: A Study of Incomes Policies in Western Europe,* Berkeley 1971.

---

income tax payments: sufficient teeth that can be enforced on a selective basis so that such teeth are not generally necessary.

[9] For example, the Internal Revenue Service is not generally construed to be a political organization nor is the Federal Reserve System. Yet, controls generally are construed as political, in part because of ideological differences and in part because the institution and administration of controls has been, thus far, a part of the political process.

Wages and the Pay Board

By DANIEL J. B. MITCHELL AND ARNOLD R. WEBER*

The Pay Board regulated wages in Phase II.[1] Although formally abolished with the onset of Phase III in January 1973, it left a bureaucratic legacy in the form of the 5.5 percent guideline and in the relatively restrained pattern of 1973 union settlements in the face of substantial inflation. When first created, the Board's labor force coverage was extensive, including virtually all private wage and salary earners, as well as state and local government and post office employees. (Wages in the unionized sector of the construction industry remained with the Construction Industry Stabilization Committee, a pre-Phase I agency.) But the Pay Board itself was small; it had 240 employees when terminated. The combination of a small staff, extensive coverage, and a strategy aimed at affecting large unionized situations dictated the shape of the Phase II policy.

Pay Board reporting requirements were designed to reduce caseload and to give the closest scrutiny to wage changes in major economic units. Units of 1,000 or more workers reported directly to the Board. The Bureau of Labor Statistics estimated that about 19 million workers fell into these units (Categories I and II) which, with the exception of certain government workers, had to report all adjustments. Smaller units of less than 1,000 workers did not report at all unless exceptions to the standard were requested. In such instances, reports went first to the Internal Revenue Service (*IRS*). Although a majority of the workforce fell into these smaller units, no more than 2 million small-unit workers had direct contact with the *IRS* or the Board.

Including *IRS* appeals, executive compensation, and nonunion construction cases, the Board handled about 17,000 cases during its life. But the reporting rules did help focus attention on Categories I and II where the major union units were located. Categories I and II were about 49 percent unionized, compared with 26 percent for the overall employed civilian labor force.

The extensive coverage also required the development of general rules which could be largely self-administered. The 5.5 percent guideline and the 7 percent exceptions to the guideline reflected this approach. Although the 5.5 percent standard was the result of a tripartite bargaining process at the Pay Board, it did incorporate a fundamental economic logic. The average wage increase for the economy as a whole should be limited to the long-term growth rate of national productivity (3 percent) plus the "target" increase in the general price level as stated by the Cost of Living Council (2 to 3 percent). It was fully recognized that some settlements would exceed 5.5 percent, but it was expected that they would be balanced by situations in which the annual wage gain was less than the standard. Thus, various "exceptions" were permitted for equitable and political reasons. Chief among these exceptions were the "tandem" rule, which permitted traditional follower units to conform to pattern setters; the "catch-up" rule, for units

---

* University of California, Los Angeles, and Carnegie-Mellon University.

[1] Space limitation requires the omission of source notes for data cited below. More detail will appear in a forthcoming volume sponsored by The Brookings Institution.

whose wages had lagged in the precontrols period; and the "essential employees" rule, for labor shortages. The last was rarely used due to the lack of strong labor demand pressure during Phase II and the stringency of the conditions established to take advantage of the exception.

I. The Influence of Tripartitism

Most of the Pay Board's policies were formulated during its "tripartite" stage, when it was composed of equal representation from (organized) labor, management, and the public. The decisions made came from a three-way negotiations process. For example, the 5.5 percent guideline applied only to "new" adjustments, those established in Phase II. "Old" adjustments established prior to Phase II were subject only to a challenge procedure, with a *de facto* 7 percent standard. "Old" adjustments were preponderantly found in the union sector under the terms of multiyear agreements. The importance of this policy can be seen from the fact that four out of five workers in major union agreements outside the construction industry were scheduled for deferred and/or escalator adjustments under old contracts in 1972. Labor was particularly sensitive to rollbacks of agreements negotiated in good faith prior to controls.

Cost-of-living escalator adjustments—which are found mainly in the union sector—were given an advantageous computation rule. (Despite this incentive, escalators declined in popularity during Phase II, according to Bureau of Labor Statistics (*BLS*) data. Evidently, the incentive was overcome by the dampening of inflationary expectations.) And other effects of the tripartite process could be cited. Still, it will be argued below that the union sector felt the brunt of the program.

II. Congressional Influence

In several important ways, Congressional action altered the design of Phase II wage controls. It reversed a Pay Board decision denying claims for about $500 million in deferred wage increases frozen by Phase I; and it mandated special treatment for the "working poor," small business, productivity-incentive plans, and "qualified" fringe benefits.

The first three special-treatment provisions probably had little effect on wage developments, although they created some administrative problems, but the last did lead to a relaxation of the standards. Qualified fringe benefits are essentially pensions, health and welfare programs, and thrift schemes. An unlikely coalition of organized labor and the insurance industry lobbied intensely for special treatment of such fringes; one of its major arguments was that qualified benefits represented saving and were therefore not inflationary. This view had apparently been persuasive during earlier wartime programs, although the relevance in a cost-push context could be questioned.

The Pay Board was forced to establish a separate standard for qualified benefits of 0.7 percent (with liberal exceptions), leaving wages and nonqualified benefits under the 5.5 percent and 7 percent guidelines. As a result, the basic standards for total compensation became 6.2 percent and 7.7 percent. Despite the shaky Congressional rationale, this flexibility probably served a useful purpose. The original price inflation target of 2 to 3 percent by the end of 1972 was not achieved, and the higher standards were more realistic.

Surprisingly, there was no upsurge of interest in fringe benefits following the modification of Pay Board standards. Commerce Department data indicate that employer contributions to pension and welfare funds have risen more rapidly than ordinary wages in the postwar period. However, during 1971–1972, the gap between the rates of growth of fringes and wages narrowed. *BLS* records on major union settlements show a drop in 1972 relative to 1971 of the percentage of workers covered by pension and welfare improvements. Some of this drop may be due to the industry composition of 1972 negotiations.[2] But the 1972 decline in the face of contrary incentives is still noteworthy. The preference for cash wages suggests that workers and union leaders were more interested in recouping lags in real disposable income than in achieving gains in deferred income during Phase II.

III. The Wage Impact

Both the published data of the *BLS* and the Pay Board's own internal records indicate that during Phase II union wages outpaced nonunion wages. The weighted average Pay Board approval in Categories I and II was 5.4—5.9 percent for union units and 4.9 percent for nonunion units. Wage pressure was most heavy in industries such as mining (8.8 percent) and transportation (6.7 percent), which are heavily unionized. Settlements in these industries include the well-publicized contracts in bituminous coal, longshoring, and railroads. In general, the larger average approvals came in *new* union settlements (6.9 percent). Seventy-seven percent of the workers in Categories I and II requested no more (and often less) than the basic standards of 6.2 percent and 7.7 percent, 75 percent in the union sector and 79 percent in the nonunion sector. Only 2 percent of nonunion workers requested over 9 percent, compared with 12 percent of union workers. Some of the higher nonunion requests were white-collar tandems to production-worker units in unionized firms.

Cutbacks occurred in about 7 percent of the Category I and II cases. Most of the reductions were in the union sector, especially for new adjustments. The average nonunion case was cut by only 0.1 percent. Old union cases (deferreds) were also cut by an average of only 0.1 percent, owing to the relative generosity of the Board's challenge procedure. For new union adjustments, however, rollbacks averaged 0.7 percent, i.e., from 7.6 percent requested to 6.9 percent approved. Of course, rollbacks do not measure the full impact of controls on new adjustments since the program presumably tended to restrain initial requests.

BLS figures also suggest that pressure came mainly from the union sector. Industries which show higher-than-average increases in average hourly earnings in 1972 tend to be those most heavily unionized. Various factors account for this. Employment and profits were expanding in the industries which experienced relatively high rates of wage increase, so that demand factors were supportive. There also was catch-up pressure on wages since the industries which experienced relatively high wage increases in 1972 appear to have lagged behind other industries in wage growth in the late 1960's. Phase II caught the tail end of a catch-up wage round, and Pay Board policies permitted the adjustment to take place. At the same time, the Board restrained the tendency to "overshoot" and thereby eliminated a possible leapfrogging process. It is interesting to note that part of the reason for the relative union restraint in 1973 seems to be an abatement—perhaps short-term—of catch-up pressure.[3]

Wage data indicate a lower rate of increase under controls than in the precontrols period. The *BLS* hourly earnings index rose by 6.9 percent in the twelve months prior to Phase I. During Phases I and II, the annual rate of increase of the index was 6.2 percent. After the post-Phase I bulge, when three months' worth of deferred adjustments was permitted to occur, the index rose at a 6.1 percent rate. Obviously, other factors might be cited to explain the slowdown. However, the reduction in union wage settlements, which were the main target for controls, is particularly dramatic. First-year wage adjustments in major agreements outside the construction industry ran at 10.9 percent in 1970, 11.6 percent in 1971, and 7.4 percent in 1972. For the first nine months of 1973, the rate has been 6.1 percent.

---

[2] Over 40 percent of the workers under major 1972 contract expirations had their original contracts written in 1969, when there was also a fringe drop.

[3] See Marvin Kosters, Kenneth Fedor, and Albert Eckstein.

Econometric attempts to disentangle the impact of wage controls from other forces tend to produce ambiguous results. First, most wage equations were heavily "massaged" in the immediate precontrol years to explain away a tendency to underpredict wage change. Thus, their reliability is highly suspect. Second, and more important, the typical wage equation is designed to explain movements in some heavily aggregated index. But it is known that despite their extensive coverage, Phase II wage controls were most heavily focused on new union adjustments. New union adjustments covered roughly one-fourth of the union workforce, which in turn constitutes only about 30 percent of private, nonfarm wage earners. Such adjustments are widely believed to play a key role in future wage patterns throughout the economy. But in the short run, even a large impact on new union adjustments has only a small effect on aggregate wage indexes. Thus, the econometric approach tends to lose impact amidst the general "noise" which characterizes all economic data series.

IV. Distortions

Wage controls might cause various types of distortions such as labor shortages. With this danger in mind, the Pay Board's economic staff went looking for shortages, but came back empty-handed. This is not conclusive proof of the absence of shortages, but other evidence generally supports this judgment. Quit rates in manufacturing averaged 2.2 percent in 1972, about the level which might be predicted on the basis of business cycle factors. Most telling of all, when controls were relaxed in Phase III, there was no sharp upsurge in wages. Indeed, the rate of increase of the hourly earnings index during the first quarter of 1973 was below the rate experienced during Phase II. The lack of labor shortages is consistent with the argument advanced in the previous section that the main impact of controls was on the union sector. Union wages are generally "above market," so that moderate restraint should not be expected to create a labor shortage, particularly in the absence of strong demand pressure.

Controls might also distort the industrial relations system. The propensity to unionize could be boosted or lowered, depending on how controls were conducted and how workers perceived them. During Phase II, however, unions won 53 percent of National Labor Relations Board representation elections, a proportion generally in keeping with previous periods. The Teamsters, who became the "insiders" at the Pay Board after the labor walkout, also experienced a win rate in line with the past.

There were practically no strikes aimed directly at controls during Phase II. About 5,100 strikes occurred in 1972, down slightly from the 1971 level and 11 percent below the 1970 figure. Strike severity (workers involved, man days lost) fell substantially—more than can be accounted for by the drop in contract expirations and by business cycle factors. Some of the relative quiet may have been due to a lowering of union wage targets. In some instances, there may have been an erosion of employer resistance. Certain employers may have decided that their optimum strategy was to agree to any demands and let the Pay Board make a cutback. This was less of a problem than some observers have suggested—the habits of bargaining contain elements of inertia. But in some instances, notably retail food, erosion was occurring.

A definite shift in contract duration did occur. Multiyear union contracts have been a notable characteristic of *U.S.* labor-management relations. However, the uncertainty created by controls appears to have induced a shortening of contract durations. In union construction, where controls were imposed prior to Phase I, this effect is most dramatic. A Bureau of National Affairs, Inc., survey indicates that in 1970, 7 percent of construction contracts were of one year's duration. In 1971, the first year of wage controls in construction, the proportion rose to 63 percent. By 1972 it hit 81 percent. The same tendency was less pronounced, but still evident, in nonconstruction: 6 percent in 1971, 15 percent in 1972. In some cases, nominal contract length was maintained, but effective length was shortened by use of wage reopener clauses. One of the factors contributing to this adjustment in contract duration, particularly in construction, was the limited incidence of escalator clauses.

V. Summary

Although Congressional directives and tripartite bargaining over policy led to anomalies in the Pay Board's program, the Board succeeded in regulating pressure on wages from catch-up and other factors. Once the major labor agreements held over from the 1971 bargaining round were resolved, wage adjustments in the unionized sector were relatively orderly and did not generate a leapfrogging process. Thus, the Board helped to restore balance to an interindustry wage structure that had been skewed by earlier economic developments. And it did so without creating severe distortions. Although the circumstances that prevailed during 1972 probably abetted the effectiveness of Phase II, the lessons learned will prove to be of value to future efforts at wage controls.

REFERENCE

M. Kosters, K. Fedor, and A. Eckstein, "Collective Bargaining Settlements and the Wage Structure," *Labor Law J.*, Aug. 1973, *24,* 517–525.

The Inflation Problem during Phase III

By Barry Bosworth*

The United States terminated its Phase II price control program in January 1973 and shifted to the largely voluntary program of Phase III. In succeeding months price inflation soared to its highest rate since the Korean War.

Analysts have put forth many differing explanations of this experience. To some, it was evidence of the inevitable breakdown that must follow the adoption of any government price control program, while others cite the premature ending of Phase II itself as a major cause of the intensified inflation. In addition, the proper direction for fiscal and monetary policy became an issue of some dispute. A shift toward a highly restrictive policy would be the normal prescription, but many observers questioned the contribution that such policies could make to slowing an inflation concentrated in agriculture, a sector with low income elasticities of demand.

In retrospect the Phase III experience offers several lessons for economic policy. The poor price performance resulted from a sharp shift in the characteristics of the underlying inflationary pressures, rather than from a breakdown of the controls. The 1973 inflation would have been severe with a Phase II system of controls, Phase III, or without controls. As a result, the past year offers very little evidence for or against the use of an incomes policy as a counterinflationary tool.

In the middle of 1971 the *U.S.* economy was experiencing a cost-push or institutionalized type of inflation unrelated to excess demand pressures. Restrictive fiscal-monetary policies, in effect for two and a half years, had been ineffective in slowing the inflation. Although falling mortgage interest rates helped hold down the consumer price index (*CPI*), none of the aggregate price and wage indices indicates a slowing of the inflation in the first half of 1971. Meanwhile, the costs in terms of idle resources became excessive.

As shown in Tables 1 and 2, during the Phase II period some progress was made in reducing the rate of inflation in the nonagricultural sector. The overall progress was less noticeable because of a growing problem with food prices. On the wage side there was a gradual but steady reduction in the size of wage settlements—particularly in the construction sector. In addition, the program made possible a more expansive fiscal-monetary policy which lowered the unemployment rate by a full percentage point.

In the first half of 1973 the rate of inflation accelerated sharply. While the transition to Phase III may have contributed, the upward pressure on prices can be traced primarily to three areas: retail food prices exploded; the international economic situation changed significantly; and several basic materials industries encountered capacity shortages which put strong upward pressure on prices. None of these three problems is the type with which wage and price controls can deal effectively.

I. Agricultural Prices

The dramatic food price increases dur-

---

* Research associate at the Brookings Institution. Andrew Carron provided considerable assistance in the preparation of this paper.

TABLE 1—RATES OF CHANGE IN SELECTED PRICE SERIES, VARIOUS PERIODS, 1970–73
(seasonally adjusted annual rate in percent)

	Pre-Control	Phase II	Phase III
	December 1970–August 1971	June 1972–December 1972	December 1972–June 1973
Consumer Price Index			
All Items[a]	4.6[c]	3.9	8.1
Food	5.0[c]	6.3	21.5
Nonfood Items[a]	4.5[c]	3.0	4.4
Commodities	3.1[c]	2.5	4.7
All Services[a,b]	7.2	3.4	4.2
Medical Care	7.1	4.4	3.7
Wholesale Price Index			
All Items	5.2	8.3	22.3
Farm Products, Processed Foods and Feeds	6.5	23.4	47.5
Consumer Finished Foods	(6.2)	(12.5)	(28.3)
Industrial Commodities[b]	4.7	2.9	12.5
Consumer Finished Goods	2.2	1.9	12.2
Producer Finished Goods	3.6	0.1	6.7
Intermediate Materials[b]	6.5	3.0	13.5
Crude Materials[b]	3.3	12.6	23.3
Private Nonfarm Deflator	4.8	2.3	4.8

Source: U.S. Bureau of Labor Statistics and Department of Commerce.

[a] Excludes mortgage interest costs.

[b] Excludes foods and feeds.

[c] Excludes auto excise tax revision.

TABLE 2—RATES OF CHANGE IN SELECTED WAGE AND PRODUCTIVITY SERIES, VARIOUS PERIODS, 1970–73
(seasonally adjusted annual rate in percent)

	Pre-Control	Phase II	Phase III
	1970:4–1971:2	1972:2–1972:4	1972:4–1973:2
Private Nonfarm Sector[a]			
Employee compensation per manhour	7.4	6.4	8.3
Excluding payroll tax increase	—	—	6.1
Output per manhour	4.8	4.6	1.7
Unit labor cost	2.5	1.8	6.4
Average hourly earnings[a]	7.7	6.7	6.5
Negotiated Wage Changes[b]			
Total benefits over life of contract	8.4	7.4	6.2
Wages only	8.1	6.2	5.3
First-year wage adjustments, all industries	10.1	6.8	5.8
Manufacturing	8.7	6.6	6.3
Contract construction	13.4	5.3	5.7

[a] Data are for nonsupervisory employees and are adjusted for interindustry shifts and for overtime in manufacturing.

[b] Average during the period. Data on wages plus benefits refer to units of more than 5,000 employees. Wage adjustments are for units of more than 1000 employees.

Sources: U.S. Bureau of Labor Statistics, *Current Wage Developments,* various issues; BLS, News Release, *Major Collective Bargaining Settlements, First 6 Months 1973,* July 27, 1973; and unpublished BLS data.

ing 1973 dominated public discussions of the inflation. The price rise resulted from several factors but the most significant was the 3.5 percent drop in world grain production in 1972 compared with an average annual growth of 3.5 percent over the previous decade. This 7 percent shortfall generated an unprecedented rise in agricultural exports of 60 percent in fiscal year 1973 with an anticipated further rise of 45 percent in 1974. Much has been written about the Russian wheat deal, but it represented only the beginning of a general surge in agricultural exports. The grain situation was made worse by the termination of fishmeal exports from Peru which led to a chaotic market for soybean products as a high protein feed substitute. The unanticipated surge of export demand had to be met by a severe reduction in carryover stocks and by price increases sufficient to reduce domestic consumption. Wholesale grain prices rose at an annual rate of 68 percent in the first half of 1973. Some details of the supply situation are shown in Table 3.

Unlike grain prices, the rising domestic price for meat products resulted only indirectly from world market conditions. Administration spokesmen have stressed demand factors; but, while of long-run significance, these have been of relatively minor importance in 1973. As shown in Table 3, the production of meat actually

TABLE 3—SELECTED STATISTICS ON PRODUCTION OF VARIOUS AGRICULTURAL PRODUCTS, VARIOUS PERIODS, 1960–74

Product and Status	Period				
Meats and poultry (inspected slaughterings; average annual rate of change in percent[a])	1960–70	1970–71	1971–72	2nd Half 1971– 2nd Half 1972	1st Half 1972– 1st Half 1973
Meats	3.4	4.7	−1.6	−1.4	−4.8
Poultry	5.2	1.1	6.2	2.9	−2.4
Wheat (millions of bushels, annual average)	1970–71[b]	1971–72[b,c]	1972–73[b,c]	1973–74[b,d]	
Production	1,351	1,618	1,545	1,727	
Total supply[e]	2,237	2,350	2,409	2,157	
Domestic disappearance	768	855	795	755	
Exports	738	632	1,184	1,150	
Ending carryover	731	863	430	253	
Feed grains[f] (millions of tons, marketing years)	1970–71	1971–72	1972–73[c]	1973–74[d]	
Production	160.1	207.7	199.8	207.7	
Total supply[e]	209.1	241.4	248.6	240.5	
Domestic disappearance	155.2	165.7	173.1	171.8	
Exports	20.7	27.3	43.1	39.7	
Ending carryover	33.2	48.4	32.4	29.0	

[a] Calculated from production data, by weight.
[b] Twelve-month marketing year, spanning the years indicated.
[c] Preliminary.
[d] Forecast.
[e] Includes imports and beginning carryover.
[f] Includes corn, oats, barley, and grain sorghum.
Sources: Survey of Current Business, March 1973, *53*, 527–28, and prior issues; U.S. Department of Agriculture, *Demand and Price Situation*, November 1973, Table 2.

declined below year-earlier levels in both the last half of 1972 and the first half of 1973. The drop in supply pushed retail meat prices in June of 1973 to 23.8 percent above year-earlier prices.

The meat supply shortages can be traced to an interaction between high feed costs and cyclical lows in cattle and hog marketings. Supplies of both cattle and hogs were also adversely affected by fears of an extensive corn blight in 1970, which encouraged marketings designed to beat anticipated higher feed prices. Price controls made the situation worse in the middle of 1973 as farmers held back on marketings in anticipation of the announced future removal of the controls.[1]

Price controls cannot deal effectively with supply shortages in the agricultural sector. But the 1973 experience highlights several serious flaws in *U.S.* agricultural policy. There has been much criticism of the Department of Agriculture for not moving quickly enough to expand production.[2] While such actions could have affected the meat situation and moderated the rise in grain prices, the magnitude of the export rise would have been very difficult to forecast accurately enough and soon enough to relieve completely the pressure on prices.

The fundamental difficulty is that the United States has followed a policy of being a supplier of last resort for the rest of the world without providing sufficient flexibility on the supply side. Government policy is currently focused upon a system of *price* supports with adjustments being made by absorbing or releasing land for production. It would be more rational to establish a stabilization fund which buys and sells commodities out of an existing stock in the short run but is required to average a zero long-run net accumulation. Such a stabilization fund would provide the short-run flexibility needed to offset the vagaries of world weather and supply rigidities, while leaving prices free to operate as an efficient allocative device in the longer run.

Going into the 1973 crop year, carryover stocks had been drawn down to very low levels. A crop failure would have had far more serious repercussions than the price rise that actually occurred. Although the risk of serious shortages was avoided by good growing weather, we will run the same risks in the coming year. Nor can we expect private institutions to hold the magnitude of carryover stocks required as insurance against world crop failures—particularly in view of the uncertainties of future government agricultural and export policies.

II. International Developments

In early 1973 the dollar underwent a second devaluation of 10 percent, bringing the total decline between August of 1971 and June of 1973 to 17 percent. The magnitude of the devaluation made it a significant influence on the overall price level even though trade is a relatively small proportion of total output. In addition, there are several reasons for thinking that the 1973 devaluation had a more substantial impact on prices than did that of 1971. In 1971 most of the industrial countries were experiencing mild recessions with a consequent reduction in the demand and prices for raw materials. This is reflected in the price decline for *U.S.* imports of crude materials between the first and third quarters of 1971. But in 1973 world prices of these commodities were either stable or rising. Furthermore, importers of intermediate and finished goods appear to have absorbed a major proportion of the 1971 devaluation (particularly in product categories constrained by formal or informal quota restrictions), but passed forward more of the second devaluation. Domestic competitors of imported goods also were free to raise their prices in 1973 (unlike 1971), which appears to have resulted in some sympathetic adjustment of domestic prices.

---

[1] For a more detailed review of the 1973 agricultural problems, see John Schnittker.

[2] Schnittker.

The measurement of the impact of the devaluation is complicated by a strong surge in industrial production throughout the world which severely strained the capacity to produce many basic raw materials. For example, industrial production of the Organization for Economic Cooperation and Development (*OECD*) countries averaged an annual growth rate of 13.3 percent between the third quarter of 1972 and the first quarter of 1973 in comparison to a 2 percent annual growth between 1969 and 1971. While it is difficult to isolate the contribution of the devaluation, prices of imported goods rose at an annual rate of 27 percent between the fourth quarter of 1972 and the second quarter of 1973. On the basis of the 1963 input-output table (adjusted for the growing relative share of imports in *GNP*), this would translate into a 1.5 percent rate of rise in the *CPI* if fully passed through.

III. Capacity Pressures

The 1973 inflation was accompanied by many reports of scarcity and capacity pressures. Some of these problems can be traced to the agricultural situation and the worldwide shortage of raw material processing capacity discussed in previous sections. These factors are particularly relevant for petroleum and petroleum-derived products, nonferrous metals, and nonfood agricultural products. But some observers claimed to see also a general pattern of excess demand or a fundamental change in the availability of basic commodities.

At the aggregate level it is difficult to substantiate such claims. The unemployment rate ranged between 4.5 and 5 percent, and wage changes showed little evidence of demand pressures in the labor market. The Federal Reserve Board index of manufacturing utilization was at 83.3 percent in the second quarter of 1973, far below the previous peaks of 1966 and 1968. The McGraw-Hill index was also below its 1966 high. Only the Wharton index at 96.7 percent indicated any general pressure on capacity.

However, some of the more disaggregate indices make clear that the capacity pressures were heavily concentrated in the basic raw materials industries. Thus, a recently released Federal Reserve Board index of utilization in major materials industries shows a rate of 94.4 percent. This is substantially above the 1966 peak of 92.0 percent and represents the highest utilization rate of the postwar period. The industries included in this index account for only 8 percent of total manufacturing value added.[3] However, they are more important in the wholesale price index because it is based on transaction weights rather than value added. Thus, these basic materials appear several times in the production chain. Items in the wholesale price index (*WPI*) approximating these industries represent 29 percent of the industrials component, but accounted directly for 62 percent of the 12.5 percent annual rise in the industrials index for the first half of 1973. There would be an additional impact as the higher materials costs were passed through to intermediate and finished goods.

The problems in these sectors reflect a variety of factors which are not simply related to the rise of aggregate demand. For many of these industries current capacity has been seriously retarded by the prior recession which was induced by policy actions aimed at controlling the inflation of that period. Thus, gains in relieving inflationary pressures during the contraction were partially offset by accelerating inflation during the expansion. Lumber provides an example of the problems caused by sharp shifts in the distribution of output. The 70 percent rise in real residential construction between the second quarter of 1970 and the fourth quarter of 1972, induced by a sharp easing of monetary policy, strained the capacity of sawmills. The result was a 21 percent rise in lumber prices in the eight months of 1971 prior to controls, 11 percent during the 14 months of Phase II, and a further rise of 21 percent in the first four months of 1973. As a result of the previous major reductions in homebuilding, sawmill capacity in 1973 was only slightly above that of 1965. The pressures on the textile industry resulted from sudden shifts in the composition rather than the magnitude of its demand. The problem of shortages has also been exacerbated by transportation bottlenecks. The huge grain shipments created a shortage of boxcars for other uses, and inefficient operation of the Eastern rail industry resulted in an accumulation of unused rail cars.

---

[3] The industries included are steel, primary aluminum, primary copper, man-made fibers, paper, paperboard, wood pulp, plywood, cement, petroleum refining, broadwoven fabrics, and yarn spinning.

The capacity shortages have been a problem of a few bottleneck industries rather than a reflection of a general level of demand. But these problems were intensified by the *rate* of economic expansion in late 1972 and early 1973. The speed with which the economy was brought back to a level of relatively high employment stands in sharp contrast to the gradual nature of the recovery of the early 1960's. It bears a greater similarity to the experience of 1955 when a rapid transition out of a recession also created bottleneck problems.

IV. Summary

The inflation of 1973 has been quite aptly labeled a commodity inflation. The rise in agricultural and import prices accounted for approximately three-fourths of the 8.1 percent annual rate of increase of consumer prices in the first half of 1973. The inflation of basic materials prices cannot be simply added to the effects of food and imports because of the problem of double counting, but the calculations of the prior section do indicate the extent to which the inflation of industrial prices was concentrated in these commodities.

In contrast to the inflation of prior years, wage costs did not play a significant role in the 1973 acceleration. As shown in Table 2, the aggregate wage indices rose at a rate equal to or below that of late 1972. However, the rise in social security taxes did provide some additional push to price increases.

At the time a great deal of attention was focused upon the effect of the shift from Phase II to Phase III. There are indications in the private nonfarm value-added deflator of Table 1 (which abstracts from imports and farm prices) and detailed indices of the *WPI* that the transition did result in an acceleration of the inflation. This is particularly true for consumer services and the finished goods indices of the *WPI* where the increases appear to be greater than would be anticipated from the pass-through of labor and materials costs. Given the dominance of other factors, however, this effect could account for but a small fraction of the inflation.

Restrictive fiscal-monetary policies would have been a very expensive means of restraining the inflation in 1973. Given the low short-run income elasticity of demand, these policies are not effective in restraining food prices. However, these problems also could not be met with a program of price and wage controls.

The inflation of 1973 differed from that of previous years and as such it required a different set of policy responses. More than anything, it represented the costs of previous mistakes. The food inflation resulted from the failure to provide a buffer stock of agricultural commodities and the inept management of the rise in exports. The magnitude of the shift in monetary policy in 1971–72 was unwise in view of the short-run supply limitations on materials for homebuilding. The bottleneck problems in the basic materials industries highlight the importance of maintaining a smooth rate of economic expansion. Sudden deviations in either direction lead to errors in planning capacity expansion and increased inflation in future periods. There does seem to be a "speed limit" on aggregate demand increases as suggested by some observers. A highly capital-intensive economy is limited in its ability to respond to short-run economic changes. Finally, devaluation, if it is to be effective, will lead to significant demand increases for export and import-competing industries. Since the devaluation cannot be anticipated, it might have been desirable to have maintained a higher level of general excess capacity at the time of devaluation in order to provide room for the adjustment. Of course, any such policy must be evaluated against the cost of increased unemployment that it implies.

REFERENCE

J. Schnittker, "The 1972–73 Food Price Spiral," *Brookings Papers on Economic Activity*, 1973, *2*, 498–507.

Price Controls in 1973: Strategies and Problems

By MARY T. HAMILTON*

Legacies of Phase II are clearly evident in the price control strategies adopted in 1973. The two basic policy concepts of the Phase II controls were cost justification and profit margin limitation. These have been retained in subsequent Phases, but with some modifications. In addition, the concept embodied in the Term Limited Pricing Agreements (*TLP*) of Phase II has become an integral part of the control mechanism in Phases III and IV. This paper examines some of the problems associated with a *TLP* policy and profit margin control.

I. Term Limited Pricing Agreements

The Term Limited Pricing Agreement was a simplified approach to price increase approvals for firms in the prenotification category. The basic concept was a tradeoff. A prenotification firm could agree to hold the weighted average of price increases to a level below the overall goal of the program.[1] In return, the firm would be relieved of the burden of filing separately for individual price increases. Such an agreement superseded all previous price approvals, and the firm relinquished the right to file for any additional increases for a twelve-month period.

The weighted average price increase had to be cost-justified in the usual manner, but cost justification was not tied to individual products. In essence, cost increases attributable to one product line could be used to justify price increases in another. The flexibility in pricing of individual products was limited to some extent by ceilings on the percentage amount of increase allowable for individual products. Compliance with the weighted average price increase was required within each quarter as well as for the fiscal year; violation could result in termination of the agreement by the Price Commission. The firm was also subject to the base period profit margin rule.

The rationale for the *TLP* policy was administrative convenience rather than economic logic. When the freeze ended, the Commission was deluged with price increase requests under "the seventy-two-hour rule."[2] The potential for reducing demands on staff resources through simplification of price approval procedures for multiproduct firms was evident. The policy evolved from negotiations in December 1971 with two major firms, one in steel and one in chemicals. The steel company requested and was granted a weighted average price increase for specific product lines with the understanding that it would not file for additional price increases for a period of one year. The chemical firm had requested approval of a weighted average price increase of 2.5 percent covering *all* product lines. The Commission directed the staff to explore the possibility of a 2 percent limit. This was accepted by the company and became the first *TLP*-type agreement, although the policy was not formally announced until January 18, 1972.

---

* Loyola University of Chicago. This paper draws on material which I developed for a forthcoming manuscript of the Brookings Institution on price controls in Phase II. The manuscript is coauthored with Robert F. Lanzillotti and R. Blaine Roberts.

[1] The goal stated by the Price Commission was to reduce the rate of inflation to 2.5 percent. The first *TLP* agreements limited the allowable weighted average price increase to 2 percent; the maximum increase was reduced in March 1972 to 1.8 percent.

[2] This rule was designed to expedite the transition between the freeze and Phase II. It permitted certain proposed price adjustments to become effective unless challenged by the Price Commission within 72 hours.

On balance, the *TLP* policy eased the administrative burden for participating firms as well as for the Commission. The advantages to the firm of filing a single *TLP* request rather than numerous requests for increases in individual product prices were substantial; however, any dollar cost savings were offset in part by monitoring expenses. The major advantages to the firm were the ability to implement price increases without delay and the flexibility afforded in relative pricing. The Price Commission staff estimated a minimum saving of thirty to thirty-five man years, or 25 to 35 percent in operational resources. On the other hand, inconsistencies among the different types of *TLP* agreements created some administrative problems for the Commission. The inconsistencies were inevitable since agreements were negotiated on a firm-by-firm basis. More than half of these agreements were negotiated in the absence of a clear set of policy rules.

Evaluating the impact of the policy on price performance is difficult. One hundred eighty-seven agreements were negotiated in Phase II; sales of the firms involved, mainly large manufacturers, were about $200 billion. *TLP*s were unattractive to services or labor-intensive firms that could cost-justify price increases well above the levels authorized by the agreements. The average *TLP* approval was 1.92 percent. In contrast, the average approval for other prenotifying firms was 5.44 percent. A simulation study made by the Price Commission indicates that on average the weighted average price increase of *TLP* firms prior to controls was larger than the *TLP* approval. This was not true, however, for a substantial number of firms. The precontrol periods used for comparison were 1969 to 1971 and from 1970 to 1971. For 86 firms, the *TLP* approval was in excess of the 1969–71 simulated increase. The 1970–71 comparison is more favorable, but still suggests that *TLP* agreements had no downward pressure on price changes for sixty-one firms. The simulation study supports the view held by the Commissioners and the staff that *TLP* approvals were too generous.

Moreover, the *TLP* approvals understate the flexibility in the weighted average of price changes. These approvals applied to increases over base prices. Increases up to base were not included. There were several instances, particularly in chemicals and aluminum, where prices were below base throughout much of 1972. Also, price *decreases* could be included in determining compliance with a *TLP* agreement. In other words, if a *TLP* firm reduced prices for some products because of competitive conditions, ability to increase prices on other products was enhanced.

The *TLP* policy, tailored to individual firms, was an administrative modification of the regulations. Alternatively, it could have been adopted as an economic policy and served as the nucleus of a control strategy. The Commission considered doing this early in 1972 at the suggestion of Arthur Burns. He proposed that firms in manufacturing might be allowed "zones of freedom" or "zones of flexibility" within which they were free to adjust prices without prenotification. The maximum flexibility in terms of the allowable weighted average price increase was to be uniform. If the limits were observed, the profit margin rule would not apply. The rationale was that investment would be stimulated or at least not inhibited. Although the

Commission rejected the proposal, the policy was an integral component of Phase III regulations.

The general rule for price increases in Phase III was that a price could be increased above base to reflect increased costs, so long as the base period profit margin was not exceeded.[3] Alternatively, a firm could increase prices by a weighted annual average of 1.5 percent to reflect increased costs. This extended the *TLP* option to all firms by establishing a "zone of freedom" of 1.5 percent. Firms complying with this zone were exempted from the profit margin limitation. No limits were set on increases of prices of individual products.

In Phase IV, the "110 percent rule" draws upon the regulations—or experience—of Phases II and III. The regulations permit prenotifying and other qualifying firms to apply a cost-justified percentage price increase to a product or service line. However, limits on individual price increases have been reinstated. The maximum price that can be charged for any item is 110 percent of the adjusted freeze price (or base, if higher), plus the weighted average price increase authorized for the product line.

The legacy of the *TLP* policy has afforded increased opportunities for flexibility in relative pricing. The economic significance is that interference with resource allocation is minimized. Individual prices can respond to demand as well as cost pressures. The inherent risk is that firms may exploit market conditions. Price increases may be concentrated on products with inelastic demand, or on those over which a firm can exercise monopolistic power. If these products have a significant weight in the price indexes, stabilization efforts will be impeded.

These possibilities led to the decision in Phase II to limit the allowable price increase for individual products covered by a *TLP* agreement. While such limits can be effective in preventing firms from implementing high visible price increases, production of some items may be curtailed if a firm is unable to recover valid dollar costs. Resolution of the problem is difficult, particularly if the rules are uniform for all firms. However, if the rules are tailored to individual firms as in Phase II, elimination of individual price maximums could be offered in exchange for a reduced (or even negative) weighted average price increase. Despite complexities of this sort, a *TLP*-type policy has merit, especially in a tight demand situation.

II. Profit Margin Limitation

The profit margin rule does not impose a ceiling on total dollar profits. It does limit profit margins. Cost-justified price increases cannot be implemented if this would result in a firm's exceeding its base period profit margin. The rule was intended to provide incentive to increase absolute profits through increased sales and productivity gains rather than through increased prices.

Some modifications have been made, but the rule generally applies to all manufacturers, retailers and wholesalers, and service organizations covered by the stabilization program. In Phase II, a firm or organization that did not increase any price above base was not subject to the profit margin constraint. This exemption was broadened in Phase III to include firms that held the weighted average of price increases to 1.5 percent or less. Phase IV regulations remove this additional leeway, but firms that do not charge a price above an adjusted freeze price have no profit margin limitation for the first fiscal year ending after August 12, 1973.

Some problems of profit margin control

---

[3] The base price was redefined as the authorized price on January 10, 1973.

are inevitable. Base period revenues, profits and so forth must be defined; a myriad of choices is available. Since the choice will affect firms diversely, inequities may result. In addition, implementation of a profit margin rule may lead to economic distortions that interfere with resource allocation.

The base period profit margin was defined initially as the weighted average of the margin in two of the three fiscal years ending prior to August 15, 1971. By January 1973, the relevance of these base years had diminished. The regulations in Phase III increased the options available by permitting the use of fiscal years ending after August 15, 1971, in determining the base margin. Since profits were low in 1971 and controlled in 1972, it seems doubtful the forward extension of the base period offers significant relief to firms experiencing low profits in the original base years. A special regulation adopted in Phase II to deal with loss/low profit situations is still in effect, but firms that do not qualify for this consideration cannot increase profit margins above base unless granted an exception.

The definition of revenues is also complex. In Phase II, the profit margin calculation excluded revenues from farming, public utilities, life insurance, and foreign subsidiaries. It included revenues from all other domestic and export sales. This was inconsistent in that prices of exports and some items sold domestically were exempted from controls. Profits from these sales were not.

In August 1972, the Commission considered excluding all export sales from both base and current period profit margins. The issue arose because the Revenue Act of 1971 included several tax changes supported by the administration. One was the tax advantage afforded by the Domestic International Sales Corporation (*DISC*) to stimulate export sales. Price Commission regulations tended to negate this advantage. An increase in profits on exports, or a shift from domestic sales to exports with a higher rate of profit, could result in a firm's exceeding its overall base period profit margin. Under these circumstances, a firm might prefer to establish a foreign manufacturing or sales subsidiary rather than a *DISC*, in order to avoid rollbacks of domestic prices. Conversely, a change in the regulations could have resulted in *de facto* violation for other exporters. If the rate of profit on foreign sales were below that of the base period, exclusion of foreign sales would raise the current profit margin—possibly above the base level. The only recourse would then be a reduction in prices of products sold in the domestic market.

Current regulations retain the inclusion of export sales of domestic firms in the profit margin calculation. On the other hand, Phase IV regulations provide relief in some instances to sellers of exempt items. If a firm during its most recent fiscal year derived 90 percent or more of its annual sales or revenues from the sales of exempt items and less than $50 million of its annual sales or revenues from sales of nonexempt items, it is not subject to the profit margin constraint.

A particularly troublesome aspect of the profit margin definition is the question of interest deductibility. The Price Commission defined the margin as net operating income before taxes and before extraordinary items. Early discussions of the Commission clearly indicate that interest, both short and long term, was considered a deductible item. The reporting forms that were subsequently issued limited the deduction to short-term interest. As a result, the profit margin limitation distorted corporate financing, not only by encouraging debt rather than equity financing, but also by encouraging shifts from long- to short-term debt.

The situation was brought to the attention of the Commission in a memorandum from the staff of the Federal Reserve Board in August 1972. Distortions in money markets were developing. The Commission felt strongly that the policy it endorsed had not been implemented, and changed the regulation to permit deductibility of long- as well as short-term interest. This ruling has been retained in Phases III and IV.

Most of the alleged distortions resulting from the profit margin rule are related to the inability to attract investment capital because of low profits. The cement industry is illustrative. In the fall of 1972, the industry claimed that the shortage of cement was attributable to lack of capacity and the inability to operate marginal plants because of expenditures required to meet pollution and safety standards. The Commission held public hearings. Testimony presented indicated that the industry had an historic record of alternating conditions of under and over supply. In view of the lead time necessary to bring new plants on stream, it seemed probable that factors other than Price Commission regulations were responsible for the difficulties in attracting investment capital. The industry's request for a class exception to the profit margin constraint was denied at that time. In the fall of 1973, the industry was exempted from controls.

These issues are illustrative rather than exhaustive. A profit margin limitation, in effect, freezes discrepancies between competitors. This may reduce inflationary pressure in competitive situations where price leaders are unable to increase prices; it can also impose burdens on small firms and possibly result in diminished competition. The profit margin constraint also discriminates against new management.

III. Summary

Price controls by definition interfere with the normal process of resource allocation through the market mechanism. Flexibility in relative pricing reduces this interference. Viewed in this context, the *TLP* aspect of controls in 1973 has economic virtue.

The economic logic of profit margin control is more questionable, particularly at the present time. Since price increases are limited to a dollar-for-dollar pass-through of costs, margin control seems redundant. Moreover, problems associated with its implementation raise doubts about its efficacy as a control strategy.

MONETARY POLICY AND SECTORAL CREDIT FLOWS

Alternatives to Monetary Policy

By JAMES S. DUESENBERRY*

The more we use monetary policy, the less satisfactory it seems. I do not say that in a spirit of criticism. Any Monday morning quarterback can tell that the plays should have been called differently. But in this case the trouble is not with the players but with the intrinsic difficulties of the game itself. Monetary policy can without doubt be a useful stabilization instrument, but its use is limited by two kinds of difficulties. First, the use of monetary policy to offset substantial variations in aggregate demand produces significant adverse side effects. Second, we know far too little about how it works. After discussing these limitations, I will consider some alternatives to monetary policy.

I. Limitations of Monetary Policy: Adverse Side-Effects

Ever since the mid-1950's, monetary restraint has always reduced substantially the rate of housing starts. I shall not attempt to reach a conclusion as to the theoretical desirability of using housing as a stabilization vehicle, if only because evidence on this question is remarkably scanty, but a number of points are worth considering. (1) The use of excise taxes to produce variations of 30 percent in a major industry's output would never be acceptable. Should monetary policy—which has similar effects on the housing industry—be immune from criticism? (2) Some proponents of monetary policy argue that, because it principally affects housing, the use of monetary policy is a cheap way to stabilize aggregate production because housing is a durable good with a short production period. But the durability of housing has nothing to do with the cost of using monetary policy. Indeed, to adapt to instability in the housing market, producers may be forced to use high-cost production methods, thus increasing the cost of using monetary policy—which in any case should be compared with the cost of using fiscal policy. (3) An additional cost to be considered is the loss to the several million people whose plans to change housing accommodation (in trading old units as well as new ones) are frustrated when mortgage credit dries up.

Those considerations do not tell us that using housing as a stabilizer for the rest of the economy is less efficient than any alternative. But points (1) and (3) do suggest that the politics of housing will inhibit the use of monetary policy as long as monetary restraint continues to bear heavily on housing.

The expansion of Federal National Mortgage Association (*FNMA*) and Federal Home Loan Bank (*FHLB*) lending activities, Government National Mortgage Association (*GNMA*) bonds, and variable-rate mortgages, and changes in the powers of thrift institutions may help to reduce the cyclical variations in the supply of mortgage credit. If these succeed in reducing the amount of rationing in the mortgage market we can expect that for any given Federal Open Market Committee (*FOMC*) action rates will be more volatile than before. Housing may still prove sensitive to monetary restraint because we expect it to have a relatively high interest elasticity. If so, the problems discussed above will persist. If not, either monetary policy will become less effective, or more drastic changes in *FOMC* action will be required to replace the expenditure restraint which has been lost.

---

* Harvard University.

In that case there will be new dilemmas for monetary policy. Outside the mortgage market the credit crunches of the past have operated largely through the peculiar nature of the certificate of deposit (*CD*) ceilings. In 1973 the financial market also had relatively large swings in short-term rates without any disturbance or disorder. Nonetheless, there is a limit to the range of variation in short rates which can be tolerated.

Business firms and financial institutions finance varying proportions of their liabilities in the short market. Some borrowers who are heavily in debt in the short market will have relatively poor credit ratings for other reasons. Very rapid increases in short rates may mean that (a) it will be difficult or impossible for some firms to earn their interest costs, (b) some lenders will trade off rate for quantity so that credit will become unavailable or available only at prohibitive cost to poor risks, and (c) some intermediaries will have to ration because they themselves are unduly dependent on a short market with rapidly rising rates. Explosive increases in rates can drive substantial firms into bankruptcy with effects which will be felt throughout the economy.

Some economists view the bankruptcy of corporations or financial institutions—even large ones—as just a part of the market process. That may be a correct view. The net cost of a few bankruptcies may be small compared to the benefits of the stabilization policy which brings them about. But there are certainly other aspects to the problem. There is always some risk of liquidity crisis and a sharp drop in investment expenditures (and if the equity market is seriously affected in consumption as well). The Penn Central case shows that good planning and fast footwork can prevent that kind of reaction. That success, however, should not lead us to conclude that there is no problem. Nor is it likely to make those responsible for monetary policy view the prospect of very rapid increases in short rates with equanimity. The fear of financial panic will always impose some inhibition on the use of monetary policy.

II. The Limitations of Monetary Policy: Uncertain Results

Though monetary policy is nowadays supposed to play a major role in stabilization policy, discussions of its effects are peculiarly lacking in precision. The response of the practitioners to the economy's chills, fevers and headaches seems to be, "Take a 2 percent increase in M growth and call me next month." Estimates of the effects of different rates of growth of M_1 M_2, nonborrowed reserves or reserves available for private nonbank deposits (*RPD*'s) differ widely in terms of the length of the response lag, the size of the total effect, and the distribution of the effects by type of expenditure.

Keynesian econometric models (e.g., James Duesenberry and Barry Bosworth, and Michael McCarthy) indicate that monetary policy works through a long chain of reactions which would do credit to Rube Goldberg. Open market actions affect bank reserves and thereby move Federal Funds rates, and at one remove, affect *CD* rates, bill rates, commercial paper rates, and presently bank lending rates. Further reactions in the flow of funds to different financial institutions then occur. These movements then influence mortgage credit flows and home building while short-rate movements ultimately influence long rates. Changes in bond yields influence business investment and also affect the stock market causing a reaction in consumer expenditures. The whole series of effects is spread out over a couple of years and even longer when multiplier and accelerator effects are taken into account. Moreover, there is uncertainty over our estimates of the parameters at each link in the chain.

I often think that the acknowledged uncertainties involved in making monetary policy the Keynesian way are the main cause of the monetarist revival. Intolerable uncertainty leads to faith in the supernatural. Jokes aside, the St. Louis equation and variants on it (e.g., Leonall Anderson and William Jordan) appear to give clear and definite estimates of the impact of monetary policy on *GNP* expenditures.

But recent tests (e.g., Stephen McNees) indicate that *ex post* forecasts outside the fitting period from the St. Louis model are significantly less accurate than similar forecasts with the simplest mechanically applied "Keynesian" model. We need not conclude from this that the Keynesian approach is better than the monetarist approach. We can, however, conclude that the monetarist approach does not seem to remove the uncertainties surrounding the use of monetary policy.

Thus, rational decisions on current monetary policy require the application of an elaborate decision tree analysis on a problem in which forecasts are uncertain, response lags long, and the effects of the control variable uncertain. An attempt to use such a rational procedure only reveals the extent of our ignorance and drives us back to small policy changes which overemphasize the current situation relative to forecasts of the future and which are typically either too little and too late or, occasionally, too drastic and even later.

III. Rules vs. Activism

Given the "built-in" stability of our economy and the limitations of our policy instruments, it can be argued that we might as well give up activist policies and fall back on monetary and fiscal policy rules. It can be argued that things would work out well enough if we maintained a balanced full-employment budget and a steady rate of growth in money supply.

While that proposal has certain attractions, a number of considerations suggest that it would not prove satisfactory. The real interest rate required to maintain a given level of unemployment with any fixed full-employment surplus or deficit is likely to vary over time for several reasons. First, the impact on *GNP* of government expenditures will depend on the mix of revenues used to finance them. Second, the balance between private saving and private investment at full employment at a given real rate of interest can easily swing as a result of technological and demographic factors by a couple of percent of *GNP*. Such a swing, if not offset, could (allowing for multiplier effects) produce a change of 2 percent in the unemployment rate.

If fiscal policy is to be fixed by a rule, then compensation for variations in the balance of savings and investment would have to come from the monetary sphere. If money supply growth is also fixed, any required adjustment in the real rate of interest would have to come from price level adjustment. If the full-employment equilibrium real rate of interest were to fall, unemployment would have to rise to induce a deflation (or reduction in the rate of inflation). Temporarily that would raise the real rate of interest, but ultimately the acceleration in the growth of the real money supply would bring about the required adjustment in the equilibrium real rate of interest. That adjustment would, however, take a rather long time and would probably produce some considerable cyclical movements.

In addition, the money supply rule itself could be a source of disturbance. We do not have a very clear idea of the rate of growth of money supply required for any one real rate of interest. And changes in the financial system must surely be expected to change money demand functions. The system would have to adjust to change in demand for money—relative to *GNP*—by changes in the rate of inflation or deflation induced by changes in employment. Again long adjustment periods and cycles would be involved.

It takes a lot of faith to believe that we can or should commit ourselves to a course determined by such a cumbersome and inefficient automatic pilot. It can, however, be argued that a gradualist approach might provide most of the benefit of a rules policy without such very long response lags. Suppose, for example, adjustments in the full-employment surplus and in the growth of monetary aggregates were made in response to current conditions and forecasts but severely limited in amount. We could then track and offset any persisting change in the demand for money, the impact of the budget, or the balance of full-employment saving and investment without any danger of overcompensating for ephemeral disturbances. But there must be exceptions to any rule, even a flexible one. There will be times when for some reason *GNP* has already moved too far away from the path of potential output, or occasions when some event makes us confident that a substantial movement away from the target path is about to occur. In those circumstances adherence to a gradualist rule will allow costly deviations to persist longer than is necessary. But it is certainly not easy in practice to choose the optimum policy. The choice will involve balancing (a) the costs of continuing excess unemployment if the economy is initially below potential output, (b) the cost of using any policy instrument to rectify the situation, (c) the cost of any "overshoot" from an expansionary action, and (d) the cost of correcting the overshoot. It is important to note that in a distributed lag system a quick recovery to potential from below must produce an overexpansion unless the direction of policy is reversed before potential output is reached. If there were no costs to the use of policy instruments, we would always correct errors as quickly as possible even though this would require—because of the distributed lag responses—frequent reversals of policy direction. In fact, every policy instrument has its costs—inefficiency in changing government expenditures, uneven allocation effects and financial instability in shifting monetary policy, and income distribution, allocational, and political costs in changing tax rates.

In a world of imperfect forecasts, long-distributed response lags to policy changes and policy instruments whose use is costly, we cannot expect to "fine tune" the economy. We must make the compromises I have already listed. Our ability to prevent or quickly correct substantial deviations from the potential output path will depend on the side costs of using our policy instruments and on the speed with which they work.

In a way the monetary policy of the whole post-accord period might be described as a gradualist policy with occasional bursts of activism—drastic moves to restraint in 1966, 1969, and 1973, and sharp reductions in short-term interest rates in recessions. The results of the activist episodes have not been encouraging. By the time that a clear case for severe restraint is recognized, it is usually too late for so cumbersome and slow moving a policy instrument as monetary policy to prevent an excessive growth in demand, the more so since action is inhibited by concern for housing and fear of financial panic, and uncertainty as to magnitude and timing of the economy's response to Federal Reserve action. In terms of the argument given above, the monetary policy instrument is too slow and too costly to provide stability against substantial shocks to the economy.

IV. Alternatives to Monetary Policy

Recognition of the limitations of monetary policy has led to a number of proposals for new policy instruments. Broadly speaking these fall into three classes: (a) variable taxes on particular types of expenditure, (b) taxes on different types of borrowing, and (c) direct allocation of credit. All of these proposals are intended to provide a substitute for monetary policy which would act more quickly, bypassing the long process of forcing up short rates and thereby inducing changes in bank lending rates and long-term market rates. At the same time they could in principle spread the costs of restraint so that housing would not be unduly affected. Finally, they involve less risk of financial shock than monetary policy.

A. Taxes

The theoretical case emerges most clearly in the cost of taxes on borrowing. In principle a tax on any particular type of borrowing should have exactly the same effect on expenditures as an increase in the interest rate on that type of borrowing (at least for those borrowers who have no other borrowing option). The difference is that the government collects instead of the lender. Superficially anyway it appears that taxes on business borrowing could be used to restrain investment expenditures while lower taxes or none were applied to mortgage loans. In practice there would be all sorts of complications. A variable tax on long-term debt issues would presumably have an impact on term structure which could dilute its expenditure impact. Mortgage funds would be used for nonhousing purposes and that might frustrate the objective. All sorts of administrative complications would arise.

Nevertheless, the basic proposition that variable taxes can substitute for monetary policy stands. Its application seems a great deal simpler in the case of expenditure taxes. Investment taxes and credits affect investment profitability in a way that is analogous to the effect of interest rates—though with a different relative effect in terms of durability. Variable taxes on consumer durables should have significant effect on timing of durable purchases. Housing investment expenditures can be taxed at the same rate as business investment or treated differently according to taste.

There is certainly no reason why the lag from the application of an expenditure tax to its effect on expenditures should be any longer than the lag from an equivalent change in short-term interest rates. To the extent that some types of investment are influenced by long-term rates rather than short, the lag for expenditure taxes should be much shorter than the lag for monetary policy. In principle, then, it appears that variable taxes and credits for investment and variable taxes on consumer durables should be able to do anything that monetary policy can do, and do it better. The intended effects on expenditure can be obtained without undesired allocational effects and without the potential for institutional insolvency or financial panic.

In practice, however, things may not work so well. The unpredictable nature of Congressional responses to tax proposals is not likely to change, and even if such variable taxes could be levied at the discretion of the President, there would inevitably be lags of several months between the time action is required and the time action is actually taken. All in all, then, variable expenditure taxes are no panacea. But in those cases when—after recognizing the lag involved—we still wish to stimulate or restrain investment, it seems to me that variable investment taxes or credits would take some of the burden off monetary policy, give a more equitable distribution of restraint as between business investment and residential construction and provide restraint without danger of financial panic. Variable taxes on consumer durables can be useful for the same reasons.

B. Direct Credit Controls

Whenever any commodity is in short supply, a popular demand for rationing can be expected. Indeed, this rule applies with special force to credit because the central bank clearly has some control over its supply. I am not at all sure that a perfectly competitive financial system would really produce an optimum pattern of resource allocation. But I also find it hard to believe that an administered system of direct credit controls—even if it were workable—would do anything but make the allocation worse. Ceilings on the amount of institutional lending to particular sectors and capital issues committees to control open market borrowing would be an invitation to evasion on the one hand and cartelization on the other. In addition, of course, the administrative burdens of a credit allocation system in a large country with a complex financial system like ours would be enormous. Most of the other countries which use credit allocation have much simpler financial systems and a far greater degree of concentration in both industry and finance. We could not follow their example if we wished.

A positive requirement that certain types of financial institutions invest given percentages of their resources in (say) residential mortgages may be workable, but would have drawbacks. Such a requirement would, of course, tend to widen the gap between the returns from mortgage lending and other investments. Indeed, the shift (by comparison with the situation in the absence of the proposed control) of (say) insurance company funds out of other markets into mortgages would push up other rates relative to mortgage rates. The result would be to weaken the competitive position of specialized mortgage lenders vis-à-vis the open market, causing a decline in mortgage lending from that source. The situation would be analogous to *FNMA* operations. And as in the case of *FNMA* operations, the regulations probably would have a net favorable effect on the supply of mortgage credit though smaller than the gross effect. But it hardly seems desirable to get snarled up in a new set of regulations to create a set of unwilling mortgage lenders. If the quantities involved were significant, lenders would be encouraged to reorganize their activities so as to move them out of the regulated sector, and other sorts of evasion would appear. If it is desired to channel funds into the mortgage market, there are lots of ways to do it, and direct allocation of private lending is not the most attractive.

Another type of allocation proposal is one made by Andrew F. Brimmer. He has suggested that differential asset reserves be applied to commercial banks. Instead of defining required reserves as percentages of different types of liabilities the reserve requirements could be defined as percentages of assets (the two could be combined of course). The reserve requirement against "good" assets, say, mortgages, could then be made lower than the requirement against business loans. The reserve ratios could of course be varied from time to time. In its economic logic this proposal comes very close to a set of taxes on different kinds of bank lending. An asset reserve requirement reduces the effective return from a loan at a given interest rate just as a tax would do. Modest differentials in favor of mortgages would have some favorable effect. However, I fear that if the differentials were large enough to have a very significant effect, they would tend to drive commercial lending out of the banking system into the commercial paper market. They might also encourage the development of new types of nonbank unregulated institutions. That tendency was visible in 1969 when the *CD* ceiling generated problems for banks. On the whole, I cannot view Brimmer's proposal with a great deal of enthusiasm.

Finally, direct regulation may be applied to the terms and conditions of lending. In the consumer credit and mortgage markets, down payment and maturity regulations have been used in the past, and they appear to be effective—particularly when applied for short periods. I see no other opportunities for that kind of regulation. We seem to have no trouble holding down mortgage credit. It does appear that if we need to do so, we can influence expenditures on durable goods by regulating the maturity and down payment terms for consumer credit. A straightforward tax on consumer durables seems preferable to me, but consumer credit regulation can provide a useful and not too costly fallback.

REFERENCES

L. Anderson and J. L. Jordan, "Monetary and Fiscal Actions: A Test of their Relative Impacts in Economic Stabilization," *Fed. Res. Bank of St. Louis Rev.*, Nov. 1968, 11–24.

J. Duesenberry and B. Bosworth, "A Flow-of-Fund Model of the U.S. Financial System," paper presented to the 1973 convention of the Allied Social Science Associations, Dec. 1973.

M. D. McCarthy, *The Wharton Quarterly Econometric Forecasting Model, Mark III,* Philadelphia 1972.

S. McNees, "A Comparison of the GNP Forecasting Accuracy of the Fair and St. Louis Econometric Models," *New Eng. Econ. Rev.*, Sept.-Oct. 1973, 29–34.

Alternative Monetary-Fiscal Policies and Sectoral Credit Flows

By Andrew F. Brimmer*

In three recent periods—1966 I–1968 IV, 1969 I–1971 IV and 1972 III–1973 IV—restrictive monetary and fiscal policies had a severely adverse impact on several sectors of the economy, particularly housing and state and local government. To assess the effect on sectoral credit flows of several alternative monetary and fiscal policies in these periods, I undertook a number of simulations using the *MIT-PENN-SSRC* econometric model (developed by economists at the Massachusetts Institute of Technology and the University of Pennsylvania with support from the Social Science Research Council and referred to hereafter as the *MPS* model). In addition to examining the effects of different monetary policies (defined as different rates of growth in the money supply), I weighed the effects of two tax policy instruments: the 7 percent investment tax credit and a 10 percent income tax surcharge.

I. Characteristics of the Simulations

To study the effects of alternative monetary and fiscal policies, the *MPS* model was simulated twice—first using actual policy (the "base simulation"), and then using a hypothetical policy.

Before these simulations were done, each behavioral equation of the model was simulated separately and the error for each equation for each quarter was recorded. To ensure that the simulated values are close to the actual values, these equation errors were then added to the behavioral equations before the system was solved in both the base and alternative policy simulations. (The *MPS* model is nonlinear; consequently, the responses to policy actions depend to some extent on the values of the simulated variables. Since the model's equations make errors, gauging the response without feeding the equation errors back in measures the response at the wrong point.)

Assumptions About the Impact of Tax Policy. In the *MPS* model, the investment tax credit works through its effect on the cost of capital. If corporations are permitted to deduct a portion of their new equipment costs from their income taxes, the effect is the same as if the price of machinery were lowered. Machinery is substituted for labor in the sense that new capacity is installed which has greater capital-output ratios than before. This in turn means larger gross investment. The extent of this effect depends on the elasticity of substitution between capital and labor. The speed of the effect depends on the length of lags between changes in the cost of capital and businessmen's recognition of it, design lags in creating equipment for the new economic factor proportions, and the lag involved in its production. Both the elasticity and the lags are difficult to measure, and doubts must necessarily remain about the degree of accuracy achieved in their measurement.

---

* Member, Board of Governors of the Federal Reserve System. This is a condensation of a paper by me, the complete version of which is available from the Board, Washington, D.C. 20551.

I am grateful to several members of the Board's staff for assistance in the preparation of this paper. Jared Enzler did the computer simulations to test alternative stabilization policies. John Austin and Ruth Robinson, my regular staff assistants, also worked on the paper at various stages. However, while I am grateful to the staff for their support, the views expressed here are my own and should not be attributed either to them or to my colleagues on the Board.

There is another effect of changes in the investment tax credit which we are *unable* to measure. This is the anticipatory effect. If the tax credit is operating (absent) and businessmen think it will be suspended (reinstated) in the near future, there is an incentive to order (hold off on) new equipment before the suspension (reinstatement) takes effect. In all that follows we assume that changes in the tax credit are unforeseen. Likewise, we have assumed that all income tax rate changes, which operate via effects on disposable income, are regarded as permanent. (This means that we have probably overstated their effects since they may be regarded as temporary).

Other Analytical Assumptions. A number of other assumptions have important bearing on the analysis. Unless otherwise stated, I assume that the policy mix does not alter (1) deposit ceilings; (2) foreign exchange rates, foreign prices, and the value of exports; (3) federal taxes and expenditures other than those specifically mentioned; (4) lending by the Federal Home Loan Banks and mortgage purchases by the Federal National Mortgage Association (*FNMA*) and the Government National Mortgage Association (*GNMA*), or (5) the Federal Reserve discount rate (although, with the money stock taken as the target monetary policy variable, the discount rate has little effect on anything except the amount of nonborrowed reserves necessary to achieve a given money stock).

Moreover, it is possible that if different policies had been followed prior to August 1971 the New Economic Policy (*NEP*) announced at that date would not have been necessary. We nevertheless assume in both the base and hypothetical policy simulations for 1969–71 that the *NEP* occurred; all policy variables except those being examined by the hypothetical policy are set at their *NEP* values. But since the model does not have a policy variable to represent the price controls instituted under *NEP*, the downward pressure of controls on the level of prices is represented by the negative residuals which occurred when the price equation was simulated separately.

There are some other effects of price controls relevant both to the 1969–71 and 1972–73 simulations which we cannot take into account. First, the controls probably altered the reaction pattern of the economic system to other policy changes. For example, in this model, the rate of change of prices depends in part on the level of economic activity. One might argue that under a system of price and wage controls, prices would be less sensitive to changes in economic activity brought about by policy shifts. While this may be true, we cannot measure it, and in the simulations reported here we assume that prices react to changes in monetary and fiscal policy in their pre-1971 manner. Second, the *MPS* model has no supply constraints except in the labor market. In the current situation, shortages are occurring partly because of capacity limitation and partly because the present system of price controls does not encourage the expansion of production of those commodities which are in short supply. In this situation, the stimulation of aggregate demand through fiscal or monetary policy may have less than the usual effect on real economic activity. This model does not reflect these developments.

The most important results of the simulation exercises are shown in Table 1 (reporting monetary and fiscal effects) and Table 2 (reporting effects on spending, prices, and unemployment). (The appendix tables to the original paper report the

TABLE 1—SIMULATION OF ALTERNATIVE STABILIZATION POLICIES, SELECTED PERIODS, 1966–1973[a]
(Monetary and Financial Effects)

Category	Year and Quarter					
	1966 I	1967 I	1969 I	1970 I	1972 III	1973 III
1. Money Stock						
Base Simulation ($ billions)	$174.6	$178.2	$205.3	$212.8	$250.8	$263.9
Index Numbers[b]						
7% investment tax credit	100.4	101.3	100.0	99.4	100.4	100.8
Improved policy mix	100.0	101.7	100.0	101.6	99.8	99.2
5% money supply growth rate	99.7	102.6	99.7	101.1	99.4	99.2
2. Interest Rates (3 mo. Treas. bill)						
Base Simulation (percent)	4.6	4.5	6.1	7.1	4.2	8.3
Index Numbers						
7% investment tax credit	93.2	85.8	100.0	110.0	92.9	91.6
Improved policy mix	98.3	75.3	100.0	80.6	104.8	98.8
5% money supply growth rate	106.2	81.2	104.8	89.7	111.9	96.4
3. Mortgage Interest Rate						
Base Simulation (percent)	6.1	6.6	7.6	8.6	7.7	8.7
Index Numbers						
7% investment tax credit	99.7	96.7	100.0	102.7	100.0	97.7
Improved policy mix	100.0	95.1	100.0	96.3	100.0	101.2
5% money supply growth rate	100.3	95.5	100.3	97.1	100.0	103.4
4. Savings and Loan Deposits						
Base Simulation ($ billions)	$111.8	$116.6	$133.7	$135.9	$200.1	$223.0
Index Numbers						
7% investment tax credit	100.1	101.2	100.0	98.0	100.1	101.2
Improved policy mix	100.0	101.6	100.0	102.6	99.9	99.0
5% money supply growth rate	99.9	102.1	99.8	102.3	99.8	98.0
5. Mortgages Held by Savings and Loan Associations						
Base Simulation ($ billions)	$112.5	$115.1	$133.8	$141.8	$197.9	$232.2
Index Numbers						
7% investment tax credit	100.1	100.8	100.0	98.8	100.1	100.8
Improved policy mix	100.0	100.9	100.0	101.4	100.0	99.2
5% money supply growth rate	99.9	101.3	99.9	101.3	99.0	98.6

[a] For more detailed data see Appendix Tables I, II, and III of the original paper.

[b] The index numbers in this table are contemporaneous index numbers with the values of the base simulation serving as the base of the index numbers. Thus, in every quarter the index number for the base simulation would be 100.0.

results in greater detail.) The figures in these tables show the difference between a control simulation (referred to as a "Base Simulation") and another specific simulation using the policy variable identified. In the first (control) simulation, all variables exogenous to the model took on their actual values. In each successive simulation, a change was effected in a given policy variable, and the responses of a number of interest rates, sectoral credit flows, and spending patterns were recorded. These derived results in turn were compared with those obtained in the control simulation. To facilitate analysis and exposition, the derived results for alternative policies were expressed in the form of index numbers where the control simulation served as the base.

II. Alternative Policies: 1966–1968

By early 1966, the economy was suffering increasingly from inflation, and the Federal Reserve shifted to a policy of monetary restraint. During the course of the year, fiscal policy also became less

TABLE 2—SIMULATION OF ALTERNATIVE STABILIZATION POLICIES, SELECTED PERIODS, 1966–1973[a]
(Effects on Spending, Prices and Unemployment)

Category	Year and Quarter 1966 I	1967 I	1969 I	1970 I	1972 III	1973 III
1. **Residential Construction Expenditures**						
Base Simulation ($ billions)	$ 27.4	$ 21.6	$ 33.1	$ 31.1	$ 54.5	$ 59.3
Index Numbers[b]						
7% investment tax credit	100.0	103.7	100.0	94.7	100.0	106.1
Improved policy mix	100.0	104.3	100.0	104.3	100.0	93.9
5% money supply growth rate	99.9	105.7	99.9	104.1	99.8	88.7
2. **Business Fixed Investment**						
Base Simulation ($ billions)	$ 78.8	$ 82.9	$ 95.5	$ 99.9	$118.3	$138.7
Index Numbers						
7% investment tax credit	100.0	97.7	100.0	101.5	100.0	97.5
Improved policy mix	99.9	96.4	100.0	100.6	100.0	95.2
5% money supply growth rate	100.0	100.8	100.0	100.3	100.0	103.1
3. ***GNP* (1958 dollars)**						
Base Simulation ($ billions)	$649.1	$666.6	$722.4	$721.2	$796.7	$841.6
Index Numbers						
7% investment tax credit	100.0	99.8	100.0	100.0	100.0	100.0
Improved policy mix	99.8	99.1	100.0	100.5	100.0	99.7
5% money supply growth rate	100.0	100.7	100.0	100.4	99.9	98.8
4. **Price Level**						
Base Simulation (*GNP* deflator)	112.4	116.2	125.6	132.9	146.4	154.9
Index Numbers						
7% investment tax credit	100.0	99.2	100.0	100.1	100.0	100.0
Improved policy mix	100.0	99.6	100.0	100.0	100.0	99.8
5% money supply growth rate	100.0	100.1	100.0	100.0	100.0	99.9
5. **Unemployment Rate**						
Base Simulation (percent)	3.9	3.8	3.4	4.2	5.6	4.8
Index Numbers						
7% investment tax credit	99.9	101.6	100.0	100.4	100.0	100.0
Improved policy mix	102.3	111.9	100.0	95.6	100.0	110.4
5% money supply growth rate	100.2	92.8	100.3	96.4	100.0	110.4

[a] For more detailed data see Appendix Tables I, II, and III of the original paper.

[b] The index numbers in this table are contemporaneous index numbers with the values of the base simulation serving as the base of the index numbers. Thus, in every quarter the index number for the base simulation would be 100.0.

expansive. As a result, economic expansion slowed noticeably in late 1966, and a modest decline occurred in real output in early 1967. Savings flows and residential construction fell sharply.

Partly in response to these developments, the 7 percent investment tax credit was suspended from late September 1966 to March 1967. An assessment of the efficacy of this policy was the subject of the first simulation. In this experiment, the tax credit was suspended on January 1, 1966, and reinstated on January 1, 1967. At the same time, the depressing effect of this action on money *GNP* was approximately offset by a less restrictive monetary policy. The results of these moves are difficult to interpret. As already indicated, in the control solution we ignore the fact that the investment tax credit was suspended for about six months beginning in late September 1966. The reason for this is that the 1966 suspension was widely anticipated and huge machinery orders occurred in September 1966. Since the suspension was relatively brief, we felt that the orders lost between October and March were about equal to what was gained by the anticipation of the suspension.

The simulation indicates that the policy would have altered the money supply, interest rates, credit flows, and the mix of *GNP* noticeably. According to the model, an increase of about $2.3 billion (or 1.3 percent) in the money stock would have been required to offset the dampening effects of the tax suspension at the point of greatest impact (i.e., 1966 IV, and 1967 I). Short-term interest rates and mortgage rates would have fallen significantly. All principal types of savings intermediaries, especially savings and loan associations (*S&L*'s), would have gained funds.

In addition, the new policy mix would have stimulated residential construction significantly. At the maximum point (1967 II), residential construction expenditures would have been nearly 5 percent above the level prevailing in the absence of a suspension of the investment tax credit, and state and local government expenditures also would have been slightly higher. But consumer spending would have remained essentially unchanged, and business fixed investment and inventories would have fallen: for instance, at the point of maximum impact (1967 III), capital outlays would have been about 3 percent below the level registered with the tax credit in force.

A second simulation experiment assessed the effects of a 10 percent income tax surcharge imposed in calendar 1966 in lieu of suspension of the investment tax credit. In this case, consumer expenditures —rather than business fixed investment— bear the brunt of the reduction in spending.

A third experiment tested still another combination of policies: suspension of the tax credit during calendar 1966, imposition of a 10 percent income tax surcharge, and reliance on monetary policy to smooth the economy's adjustment to fiscal actions. These joint tax measures go a substantial distance toward evening out savings flows in this period. Nevertheless, they offset only about one-quarter of the decline in residential construction that actually occurred. Moreover, mortgage rates still rose sharply in the simulation.

The economy emerged from the 1966–1968 period with an unemployment rate of about 3.5 percent and continuing strong inflationary pressures. In the light of this experience, we sought what we thought was a better policy mix for the period. In this exercise, the investment tax credit was suspended, and the income tax surcharge was imposed in calendar 1966, with the dampening effects on *GNP* being offset only partially by expansive monetary policy. In this exercise, residential construction drops sharply, though not by as much as the drop which actually developed. Savings flows are increased, interest rates are reduced, and the unemployment rate ends up just under 4.5 percent. Most importantly, the inflation rate ends up about 4.5 percent, about 1.5 percentage points below the actual level which was recorded. The simulation results suggest that this combination of stabilization policies certainly would have put us in better shape going into 1969–70.

Steady Money Growth. Some simulation experiments assumed steady growth in the money supply. The tables suggest that the effects of a constant 5 percent money stock expansion beginning in the first quarter of 1966 would not have been appropriate for that period: a money supply greater than the one actually prevailing in 1966, lower interest and mortgage rates, and sizeable increases in the inflow of funds and mortgage holdings of savings intermediaries. Residential construction and other sectors would have expanded strongly and the unemployment rate would have been pushed to an exceptionally low level. The net result would have been a sharp acceleration in the rate of inflation.

The results are much better when the money supply grows at 4 percent. This is fairly close to actual policy through 1967 and leads to slightly lower unemployment rates for 1966–67 than actually prevailed. However, in 1968 real growth speeded up, largely accommodated by high money growth rates and moderate interest rate increases. The hypothetical 4 percent growth rate drives interest rates very high toward the end of 1967. All in all, a 4 percent rate of growth in the money supply, combined with Regulation *Q* increases (ruled out in the simulations), would have worked fairly well. So, in the light of these results, the increase in money growth actually registered in 1968 was probably a mistake—as a number of Federal Reserve spokesmen have acknowledged.

III. Alternative Policies: 1969–1971

In 1969–71 actual fiscal policy differed significantly from that prevailing during 1966–68. For instance, the investment tax credit was suspended in April 1969, and reinstated in August 1971; and a 10 percent income tax surcharge took effect as of January 1, 1969, was reduced to 5 percent in the first half of 1970 and was then eliminated entirely. What would have been the effect of a decision to retain the tax credit and to eschew the surtax?

The first simulation of the 1969–71 experience assumed that the investment tax credit was retained, with its expansionary effects offset by more restrictive monetary policy. After a noticeable time lag, business fixed investment expands appreciably, about 3 percent more than the increase which would have occurred had the tax credit been suspended.

The reduction in the money stock necessary to offset the effect of this increase in investment on the level of income causes mortgage and interest rates to rise sharply. For example, at the point of maximum impact of this pressure on the money and capital markets, short-term interest rates would have been one-fifth higher than those which actually emerged.

Reflecting these developments, savings flows to—and mortgages held by—*S&L*'s and other financial intermediaries would have dropped sharply, imposing a severe drag on residential construction. By the fourth quarter of 1970, the margin between actual and anticipated residential construction spending was as much as $2.7 billion. But business fixed investment would have been about $3 billion higher—thus considerably exceeding the short-fall in residential construction expenditures.

On the other hand, if the 10 percent income tax surcharge had been ended completely on January 1, 1969, the differential effects on credit flows to the housing sector would have been substantially adverse, and residential construction would have suffered accordingly. While varying widely, short-term interest rates would, on balance, have increased over the period, and mortgage rates would have risen steadily. In the face of Regulation *Q* rate ceilings, this would have depressed inflows to thrift institutions, and thus residential construction as well. In contrast, outlays for business fixed investment would have been dampened only slightly. Consumer spending would have shown little or no adverse effects.

In the next experiment, the two fiscal policy measures were combined—that is, as in the first experiment, the investment tax credit was retained, and also the income tax surcharge was ended—with a more restrictive monetary policy offsetting the expansionary effects of these fiscal changes. This combination makes the situation much worse. Both short-term yields and mortgage interest rates increase sharply. Savings flows decrease, net mortgage acquisitions are substantially lower, and residential construction expenditures fall dramatically. On the other hand, both consumer spending and business fixed investment rise to higher levels.

We next considered whether some combination of policies could have brought the economy through the 1969–71 period more successfully than did the mix actually used. This time we did not change tax policy from that which actually occurred. Instead, enough money was injected into the banking system to make the unemployment rate fall to about 4.5 percent. This course substantially increases credit flows and residential construction. However, this improved sectoral performance is purchased at the price of a significant quickening in inflation: the *GNP* deflator ends up at 1 percent above the level it otherwise would have reached. On the basis of these simulations, it appears that we cannot devise a policy nearly as satisfactory as the one we devised for 1966–68. By 1969, the rate of inflation was higher than in 1966, and the steps necessary to reduce it would have been much more painful. Even the very firm policies followed in 1969 did not get the inflation rate below 5 percent in 1971 until the *NEP* began in August of that year. This result provides good evidence as to why in reality stabilization policy should not be viewed as a series of isolated episodes.

Money Supply Growth. The final simulation for the 1969–71 period involved experimentation with a constant money supply growth rate. By the beginning of 1969, the inflation rate was high enough so that a 4 percent money growth rate would no longer provide a satisfactory level of real money balances. So a steady 5 percent money growth rate was tried beginning January 1969. This policy too works out fairly well. It causes slightly lower unemployment rates than actually occurred and very slightly higher inflation rates. Residential construction and savings flows are in better shape in early 1970 than they were under actual policy.

IV. Alternative Policies: 1972–1973

As 1972 unfolded, it became increasingly evident that the economy was expanding too rapidly. Given this outlook, the Federal Reserve adopted and later intensified a more restrictive monetary policy, effected principally through open market operations and discount rate increases.

The first simulation for 1972–73 suspended the 7 percent tax credit in mid-1972 for a period of four quarters with the adverse impact on *GNP* being offset by monetary policy. As expected, this move would have resulted in somewhat lower interest rates—including those on mortgages, and thus, in this more favorable environment, residential construction would have been stimulated appreciably. At the end of 1972 (partly because of the long lead times involved in construction activity), the net increase would have been rather slight (only about $200 million). However, for 1973, the additional spending would have been significant—amounting to $4.3 billion as the year ended. On the other hand, consumer expenditures would have been affected very little. Business fixed investment would have dropped substantially—being $3.8 billion lower by the end of 1973.

The next experiment assumed imposition of a 10 percent income tax surcharge over the period mid-1972 through mid-1973. In this case, interest rates would have generally eased off somewhat below the levels actually reached during 1972 and 1973. Savings flows and mortgage acquisitions would have expanded noticeably, and residential construction would have received a strong boost. Consumer expenditures would have changed very little, and business spending for equipment and structures would have been somewhat higher. This experiment was followed by a simulation in which both taxes were altered, and—to the extent feasible—the dampening effects were offset by monetary policy. However, the resulting rate of growth in the money stock required to achieve this result was particularly high. In this example, interest rates were reduced considerably, and savings inflows at thrift institutions were strengthened measurably. Mortgage acquisitions and construction activity were correspondingly much stronger. Again, however, the consumer sector was little affected.

For this episode also an attempt was made to find a better policy mix. In this case, the investment tax credit was suspended for a year, and a somewhat more restrictive monetary policy was imposed in July 1972. Under this policy the inflation rate is reduced by about one-half percentage point. The unemployment rate does not go below 5 percent. Savings flows and residential construction apparently would have undergone somewhat sharper declines than those which actually occurred.

Money Supply Growth. Finally, a simulation was run for the 1972–1973 period using a 5 percent constant growth rate for the money supply. However, by 1972, the inflationary problem had become sufficiently severe that some observers could question whether the 5 percent money growth policy was any longer feasible. In the simulation reported here, this policy is sufficient to reduce *GNP* by $20 billion and to add .6 percent to the unemployment rate by the end of 1973. It also reduces residential construction by $7 billion. This is about as restrictive a policy as I personally can believe could have been carried out over the period.

Thus, we see that, as we move chronologically through the periods analyzed here, the inflation rate is successively higher. For example, although not shown in the quarterly data, the *GNP* implicit price deflator increased at an average rate of 3.8 percent in the 1966–68 period; by 5.0 percent during the years 1969–71, and by approximately 6.8 percent during 1973. Because of this trend, any constant money growth rate appears to become a progressively restrictive policy in each time period. Thus, on the basis of the simulation experiments reported here, a 5 percent rate was too expansionary in 1966; about right in 1969, and apparently too restrictive in 1972–73, given the underlying price pressures.

These results raise a difficult question: Given the high rate of inflation that had already become imbedded in the economy even before the oil crisis gave prices a further boost, should the Federal Reserve have accommodated some of this inflation with money growth in excess of the rate consistent with approximate price stability? Some observers who answer affirmatively stress that—with a rate of increase in prices as high or nearly as high as the rate of growth of money—it is vital to ensure that the real money balances created are sufficient to allow for normal economic growth. For them, the best policy is to set a money growth rate not too far below that necessary to accommodate the higher rate of inflation and then to reduce successively the money growth rates later—once the inflation rate begins to fall.

This suggestion, however, is immediately subject to the criticism that—since the rate of inflation depends (with a lag) on the rate of growth of money—such a policy could set off a price-money-price spiral in which increased inflation was met with increased money growth—followed by still further increased inflation.

Clearly, the Federal Reserve must be careful not to fall into this trap. Yet, both of these lines of argument beg the question as to just what money growth rate is consistent with what rate of inflation. That is a question to which we have not so far devised a very reliable answer.

V. Summary and Conclusions

Three main conclusions emerge from the simulations. First, a comprehensive assessment of stabilization policies during the last decade strongly suggests that the economy's sectoral performance could have been improved—if such heavy reliance had not been placed on monetary restraint in the fight against inflation. Moreover, residential construction would have carried less (and the corporate sector more) of the burden of restraint if the mix of stabilization policies had been different.

Second, a hypothetical policy can be devised which gives more satisfactory results for 1966–68 than did actual policy; but by 1969–70 the pace of inflation had unfortunately progressed to the point that a policy simulation as satisfactory as the one for 1966–68 could not be constructed. For 1973 the problem becomes somewhat more difficult still.

Finally, the simulations indicate that a policy of constant growth in the money supply, if it does not accommodate to some extent the established rate of inflation, will have seriously adverse effects on the economy.

MAJOR ECONOMIC PROBLEMS OF THE 1970's

The Clouded Crystal Ball

By ROBERT L. HEILBRONER*

I shall begin this brief discussion of the policy issues of the 1970's by taking a look backward, rather than forward, reviewing some of the major issues of the future I should have raised were I speaking in the early 1950's instead of the early 1970's.

With perfect foresight, I would, of course, have mentioned the issue of economic growth, clearly visible as the central thrust of all the advanced industrial societies; the recalcitrant problem of inflation; the constraint of the environment as a wholly new and important element for both micro and macro policy; the difficulties implicit in Japan's remarkable reentry into the world economy; the emergence of the multinational corporation as a new agency for the conduct of international economic relationships; the problem of economic development, stubbornly resisting attempts to move traditional societies off dead center by the application of standard economic nostrums; and the problem of the dollar and what to do about it.

Three aspects are common to all these "future" problems of the past. First, these are unmistakably economic, rather than political or sociological. Second, they are all deep-seated rather than "accidental" or superficial economic problems. Last, and most important, *every one of these problems was invisible in the 1950's*. In point of fact, had I really been holding forth on the outlook in the early 1950's, I doubt that I would have mentioned a single one of them.

For in those earlier years, it was not growth but the threat of chronic recession that still absorbed the attention of the profession. Inflation was a matter on which no sessions were organized, because we knew that it could not occur as long as unemployment was 4 or 5 percent of the labor force. The Sierra Club may have been exercised over the deterioration of the environment, but not the American Economic Association. No one had heard of the multinational corporation. Japan was considered by all to be a hopeless economic invalid. The prevailing attitude toward economic growth in the less-developed countries was one of encouragement, not to say enthusiasm: I believe the 1950's were officially dubbed the Decade of Development. The problem of the dollar, you will remember, was that it was deemed to be in more or less permanent short supply.

Now I do not mention this extraordinary myopia to take the profession to task. I raise it, rather, to inject a cautionary note in making our projections and predictions today. Like everyone else, I have my list of expected policy issues of the 1970's—a "surprise-free" list, in Herman Kahn's terminology. It includes the very problems I have just enumerated: growth, inflation, the environment, the multinationals, the failure of development, the international monetary situation. But I am moved to ask, reflecting on the past, whether this list of problems is likely to be as misconceived as an earlier list would have been; and what unexpected, truly surprise-laden, issues should be included on that list.

---

* Norman Thomas Professor, New School for Social Research.

Before I attempt to peer into the clouded crystal ball, let me ruminate on why the great body of economists failed to predict the major trends of economic affairs over the past two decades. I can suggest four reasons for this failure: perhaps there are others. First, the perceptual capabilities of the human mind make it genuinely difficult to perceive new "problems" in any field. Events bombard us in seemingly random fashion. We are assaulted by the present. There is no way of gathering, or having gathered, of rank-ordering in importance, the data of contemporary existence. Thus we rarely discern a problem that is "latent," although after the fact it is all too easy to discover its premonitory warnings.

The second reason is related to the first, yet distinct from it. It is that we tend to organize our perceptions according to received doctrines which are well understood—and find it difficult to formulate new paradigms that will reorder data according to a different set of rules. For example, economists struggled first to accept Keynesian ideas and then to move beyond an oversimplified Keynesianism; Marxist economists struggle to escape from the mechanical predictions of vulgar Marxism.

Third, it is difficult to formulate cogent long-term trends because the indeterminacy of the economic system is greater than in the past. The policy issues of the 1970's will not emerge solely as the outcome of blind market forces, but will be "made" in Washington, Tokyo, Paris, Bonn. One can no longer assume that a LaPlace-like intelligence, equipped with a knowledge of the Walrasian equations, will suffice to foretell the future. George Shultz is apt to get in the way.

Fourth and last, predictions are difficult because there exist connections between economic trends and sociological or political behavior patterns about which we know very little. At some level, impossible to tell in advance, affluence begins to change patterns of consumption and of work. At some level, equally impossible to foretell, thresholds of social tolerance are overstepped by economic events, and backwash effects take over, self-fulfilling prophecies begin to dominate economic activities, expectations change permanently. In turn these changed social and political parameters alter the economic variables. But not knowing in advance what the "cross-elasticities" of sociopolitical and economic behavior may be, we cannot foretell when the rules of the game will be changed.

With these caveats, it would be a foolhardy man who would presume to suggest issues for the 1970's, other than those that we all expect to dominate the stage. Let me nonetheless suggest a few additions to the surprise-free list:

1. *Increasing micro-breakdown*

I mean by this that a major problem of the 1970's may be severe dysfunction, or even nonfunction, of *parts* of the economic system, rather than with its main macro-aggregates. As straws in the wind, and perhaps examples of the future, let me enumerate the near-thing on Wall Street, the debacle of Penn-Central, the almost bankruptcy of Lockheed, work discontent in high paid factories, the crisis of the cities. I think we must expect more of this in the future.

For what reasons? Partly, I would suggest, because a new macrostructure "permits" micro-instability to manifest itself without the self-corrective cure of bankruptcy, massive unemployment, etc. Partly the reasons may lie in new attitudes at entrepreneurial and worker levels: a decline in the "work ethic," as the current saying goes. Partly the cause may reside in a bureaucratization of public and private institutions that leads to a deadened capacity to respond.

2. *Increased tension over income distribution*

Why social peace persists in the face of an outlandishly skewed distribution of income is a perennial puzzle, especially in a democratic society. Why do so many acquiesce in an income distribution that accords so much to so few?

We do not fully understand the reasons, but among them is certainly the parochial range of reference groups. Policemen compare their earnings with firemen, not with Wall Street lawyers. This is because they recognize the "legitimate" superiority of Wall Street lawyers—evidenced in the fact that they make so much more money than policemen!

I suggest that this tolerance for inequality may be waning. The reference groups against which the middle groups of the population measure the "fairness" of their rewards are widening. This is vividly illustrated in the "exorbitant" demands of certain strategic unions, or the jockeying for parity among municipal employees who were formerly resigned to a certain pecking order. To the extent that environmental or other difficulties require a slowdown in the pace of material economic growth, this struggle for place is apt to become more marked. What was formerly an orderly procession, in which each knew his approximate place, degenerates into a free-for-all. The results would be not alone an intensification of inflationary pressures, but a reappearance of that "vanished" problem: Disraeli's War Between Two Nations—the Rich and the Poor.

3. *The decline of neoclassical economic theory*

This is hardly a problem for policy, but it is a major problem for policy makers. Theory is the way in which we try to arrange the facts of life, or the guideline by which we seek to rearrange them. Much, although not all, of contemporary theory is of almost negligible importance for policy purposes. This is because it is built on assumptions with regard to behavior—both individual and institutional behavior—that are too far from reality to produce models in which we can place much confidence. I once said (and was properly reprimanded for saying) that if 99 percent of all economic theory disappeared, leaving us with little more than a freshman knowledge of markets and of the main macrovariables, our ability to conduct economic policy would not be one whit affected. If I am correct in anticipating massive micro problems and mounting tensions over income distribution—in addition to the standard assortment of ills to which the economic body is heir—we will need a considerable advance in theory if we are to deal with these problems other than on a catch-as-catch-can basis.

There are two directions in which theory needs enlargement or renewal. One is a theory of institutional behavior, the counterpart, so to speak, of Marshall's theory of the representative firm. We need to know more about the representative institution: how it receives and transmits information, how it comes to decisions, what criteria (certainly not "maximizing"!) it uses for choice, how economic power is exerted on it and by it. J.K. Galbraith has attempted such a theory, on the basis of which he has erected a bold, but consistent, program for policy. If one disagrees with that policy, the need is to improve on the theory of institutional action that underpins it.

Second, we need a better theory of economic policy itself. Beginnings have been made by Adolph Lowe, Jan Tinbergen, Kenneth Boulding and a few others. Here the main effort lies in the effort to rethink the relationship between variables and parameters. In a world of public intervention, does orthodox economic behavior remain a constant on which policy prescriptions can be reliably based? Or does economic behavior—individual or institutional—become the target variable, the object of manipulation in order to attain desired goals? This raises a host of questions. By what means can one influence behavior? With what degree of confidence? To what lengths? For what ends?

I do not have policy prescriptions for the issues I have raised. I suspect that if microdisorder, an intensified distributional struggle, and a growing awareness of intellectual incompetence do in fact become major issues for the 1970's, we will meet them, as we have met other new issues, by *ad hoc* responses and a gradual reformation of ideas. But of course the real issues of the 1970's may be something quite different from what I have suggested. Herodotus declared that the only thing one could say about the divine influences that direct the course of history was that they rejoice in disturbing and upsetting things. He must have meant economic history too.

Some Conjectures on Policy Problems of the 1970's

By KERMIT GORDON*

I understand this program to be an exercise not in prescription but in prediction, a much more hazardous form of intellectual exposure. A quick way to acquire a suitable modesty about our capacity to see clearly ten years ahead is to review the state of the predictive art a decade ago. The effects are sobering.

Ten years ago some social thinkers—but I must say to be fair, not many economists—had discerned a new menace to social peace called automation. Somehow this newly discovered force was destroying jobs so fast that work itself was in danger of becoming obsolete, and we faced the challenge of designing a new society in which work would figure only incidentally. It is embarrassing today to recall that automation would have drawn many votes in 1963 as one of the major economic problems of the 1960's.

With this cautionary example in mind, I approach with humility and diffidence the task of identifying some of the major economic policy problems that are likely to absorb our attention through the remainder of this decade.

International Economic Problems. We live in an increasingly interdependent world in which the institutions and arrangements that nurtured and sheltered this interdependence have broken down. World trade has been growing much faster than world income; international investment, whether via the multinational corporation or otherwise, has expanded steadily; money and capital markets of the industrialized countries have become integrated to an unprecedented degree. But at the end of the era during which these links were being forged, we saw the collapse of an international monetary system based on gold, dollars, and fixed exchange rates, and we saw large fissures appear in a trading system that was characterized by increasing liberalization and nondiscrimination and that was shaped under American leadership and lubricated by *U.S.* balance-of-payments deficits.

The present situation is clearly unstable. The United States has lost the dominance that enabled it to exercise effective leadership in the creation of new institutions and arrangements, and other sources of leadership have not yet appeared. The increasing fragility of the global economy comes at a time when both North-South economic relations and East-West economic relations are in flux. Though the developing countries as a group have enjoyed a high rate of economic growth, they are troubled by problems of rapidly rising population, unemployment and maldistribution of income. The grievances they feel against the industrialized countries are intensified by a rising spirit of nationalism. Clearly the donor-client attitudes which characterized the foreign aid relationship in years past have come to an end. The developing countries are convinced that the terms of trade have been rigged against them, and they can be expected to use whatever means are available to improve their relative price position.

At the same time, the Communist countries are seeking more intimate trade and credit relationships with the rest of the world. This development will raise again the complex of questions concerning the mechanisms of economic intercourse between state trading countries and capitalist countries that have been discussed since the end of the Second World War but never resolved.

---

* President, The Brookings Institution.

Inflation. Some economists squirm a bit when they hear themselves characterizing inflation as one of our gravest economic ills. They probably have in mind a hypothetical society in which all money claims are indexed to the cost of living, and they suspect that life in such an economy might be quite tolerable. But whatever life might be like in such an imaginary society, inflation is clearly perceived by the American people as a major social disorder. When a sharp rise in the price level coincides with rising real per capita consumption, it must follow that many people are better off than they were before; yet the evidence seems to indicate that most people *feel* worse off, including many of those who are demonstrably better off. It is dangerous to dismiss as irrationally grounded the socially corrosive effects of inflation.

We should take small comfort from the fact that in recent years we have one of the better price records among the industrialized countries. In recent years we have had, on American standards, an unacceptable rate of inflation in recession, in recovery, and in expansion—a process complicated by two dollar devaluations, a world food shortage, and an energy crisis. I know of no neat theory of inflation that fits the facts of the last five years—neither aggregate demand, nor money supply, nor labor power, nor oligopoly power, nor bottlenecks, nor expectations—though I could easily be convinced that all played a part.

Until someone discovers the Rosetta stone that unlocks the mysteries of the inflationary process and prescribes the appropriate aggregate or structural measures to combat it, I suspect we will continue to see the intermittent use of the blunt device of direct governmental intervention in private wage and price decision making. I for one am not prepared to write off such interventions as inherently mischievous. A great deal depends on the temperature of the economy at the time the intervention occurs, the scope of the program, the skill and integrity with which the intervention is managed, and most of all, on the moral authority of the governmental intervenor. But however we cope or fail to cope, it seems likely that the inflation problem will remain high on the American agenda in the years ahead.

Performance of the Public Sector. Economists at all points in the opinion spectrum can rally round the proposition that government at all levels ought to be performing much better than it is. The main but not the only impetus to this convergence of opinion was, as we all know, the pervasive disappointment with the results of the new social programs of the 1960's.

In retrospect, we see now that the federal government of the 1960's attacked the complex social problems of the era with the implicit conviction that common sense or simple deductive reasoning was sufficient to design public programs that would cure social ills. Only when it was recognized that the disbursement of large sums of money does not necessarily solve problems did attention begin to turn toward the search for strategies and techniques that work.

This search for ways to translate public dollars into visible social gains will, I think, be a principal concern of public policy thinking throughout the 1970's and beyond. Let me give just one example, albeit a very important one. To move people to act in a desired fashion, one can either order them to do so or make it worth their while to do so. That is, one can choose between administrative regulation on the one hand and a system of incentives or disincentives operating through the market on the other. There is a place in the governmental system for each strategy, but in most cases we seem to go the regulation route without weighing carefully the pros and cons of each approach.

The search for effective strategies to treat social ills will rarely lend itself to armchair analysis. Often the only way to find out what works is to try out alternative strategies on an experimental basis. The social experimentation approach, though difficult, expensive, and time-consuming, deserves to be a major element in the effort to improve the performance of the public sector.

The problems are not going to evaporate, nor is government going to get smaller. It simply has to do better.

Distributive Equity. The quest for greater equality of opportunity and the quest for reduction in the inequality of results are closely related but are not, of course, synonymous.

The problem of improving equality of opportunity relates mainly, but not wholly, to minority groups and women. It is a safe bet that minority rights and women's rights will continue to occupy a central position on the national agenda of economic problems throughout the 1970's. The two movements went through an unsteady initial period, but both have clearly settled down to an effective strategy of skillful and persistent application of pressure combined with appeals to conscience.

The broader problem, the distribution of income and wealth throughout the economy, is more diffuse and ambiguous. For my part, I doubt that the present degree of inequality of income and wealth is necessary to safeguard the motivations and incentives which drive the system. Some forces are at work to moderate the degree of income inequality, such as the minority and women's movements and the great expansion of government transfer payments and subsidized services to low-income groups. But earlier visions that saw the spread of unionization and collective bargaining and of higher education as equalizing forces are held today with considerably less fervor.

The tax system is potentially a powerful device for reducing nonfunctional inequalities in income and wealth. Yet the total American tax system today does not appear significantly to alter the pretax distribution of income, nor does our present system of death duties seem to have much effect on the distribution of wealth. Given the inertia of our tax system, it would require a major crystallization of public opinion to achieve changes in the tax system that would serve even moderate redistributive ends. My tea leaves do not tell me whether such a crystallization is in prospect in this decade.

Environment, Energy, Resource Depletion, and Economic Growth. Though these are by no means a single problem, they are linked together in a network of interconnections. We are now well beyond that rapturous early period during which some apparently felt that purity of heart and worthiness of motives were capable of cleaning up the environment. As the problems have shaken down, it has become clear that environmental progress involves painful choices, complex trade-off problems, and a critical evaluation of fundamental strategies and techniques. Despite occasional and even major setbacks to the cause, the environmental issue probably possesses staying power: any movement grounded in an alliance of youthful activists and middle-class conservationists is likely to continue to make itself heard.

The long-term dimensions of the problems of energy and resource depletion are not nearly so clear. If only free markets ruled the world, it would be possible without great difficulty to visualize the kind of economic and technological measures—working through economizing, substitution, and innovation—that would be required for a smooth adaptation to the gradual depletion of particular wasting resources. But the world is also influenced by monopolies, international politics and ideologies, and military force. The crisis we are now passing through, however it may be resolved, is bound to alter fundamentally the way the United States and the rest of the world think about long-term energy and wasting-resource problems. True, there is probably no internationally traded commodity other than oil in which the potential for coercion and gouging is as great, but there are probably a few other commodities the monopolistic control of which is capable of causing significant disruption. Many of us would have said even as recently as a year ago that measures to achieve the goal of national self-sufficiency in energy sources would cost us dearly and would do great harm to the world trading system. Today it is much more difficult to build a case for a free trade policy in oil.

Any forecast made in the midst of an imposed world oil crisis is almost certain to be distorted by short-term perceptions. But it takes no great acumen to say that the policy problems associated with energy sources and wasting resources will be with us in one form or another in the years ahead and will occupy a position high on the economic policy agenda.

For the sake of completeness, it is necessary to take note of the view of those who would halt economic growth as a means of saving the natural environment and slowing the depletion of resources. This could be a prescription, not for throwing out the baby with the bath water, but for throwing out the baby instead of the bath water. But whatever the merits, the no-growth debate is likely to continue to stimulate analysis and arouse passions throughout the decade.

Economists as Servants of Power

By SAMUEL BOWLES*

Radicals are often asked, "What are you for? We know what you are against." Hardly an unfair question, particularly in a discussion of policy problems in the 1970's. But the answer, it turns out, casts doubt on the value of discussions such as this, concentrating as they do on policy alternatives for the state within a capitalist framework. What we want is equality and democracy. So do you. By equality we mean the absence of involuntary distinctions of wealth or consumption and the obliteration of racially or sexually-determined social roles. By democracy we mean a process by which people can exert a substantial influence on those institutions—the polity, the family, the workplace, the community—which govern both our own individual self-development and the evolution of our society as a whole.

These objectives, you will note, conflict directly with the perpetuation of those institutions which define the capitalist system. Worse still, a worker and citizen-controlled egalitarian society is not merely unattainable under capitalism, but in the contemporary United States is beyond the pale of responsible political discussion. In short, most of the issues which concern the radical economist are not "policy issues."

The severely constricted range of policy alternatives in capitalist society has not dampened conventional economists' ardor for involvement in the state. In fact, for a major group of economists, whom I will call the "Engineers," policy and relevance mean the same thing. Their enlightened dedication to serving the state reflects, no doubt, a secular version of the faith of one of the founders of the American Economic Association. "God works through the State in carrying cut His purpose more universally than through any other institution," wrote Richard T. Ely (cf. Sidney Fine, p. 180). To the other major group of economists, whom John Kenneth Galbraith calls the "Priests," neither policy nor relevance appears to mean anything. Yet in their social function (if not intent) these two groups of economists are highly complementary. The Engineers—neo-Keynesian stabilizers, public finance economists, human resource types and the like—figure out ways to ameliorate social conditions which run the risk of becoming politically explosive. Other Engineers work to "rationalize" production and increase profits for the large corporation. The Priests—neoclassical theorists for the most part—are less useful, but no less important. The impact of their work is ideological: to obfuscate the sources of social problems, locating the roots of inequality and hierarchy in the nature of "man" ("preferences") or the state of nature ("technology"), and thereby denying the possibility of a substantially better society. (Critiques of the work of the neoclassical Priests and the neo-Keynesian Engineers can be found in Edward Herman and Richard DuBoff, E. K. Hunt, and Herbert Gintis.) In their role as apologists for the capitalist order, as much as in their role as advisors to decision makers with control over substantial blocs of resources, economists have become the new servants of power.

This dubious achievement has its roots in the rise of the modern corporation and the extension of the state involvement in

---

* Harvard University and University of Massachusetts at Amherst.

social life beginning at least a century ago. The political mobilization of the working class and discontented farmers in the last quarter of the nineteenth century and the rise of popular-based urban political machines posed a dilemma to the ruling class. An extension of state activity in the interests of amelioration of explosive social conditions, repression of dissident groups, and coordination of an increasingly interdependent economy brought with it the possibility of this newly empowered state falling under popular control and ultimately being used against the capitalist elite and its allies in the professional classes.

The modern liberal solution was to welcome the extension of state power and simultaneously to insulate the implementation of state action from popular control. The ideological justification for this process took the form of what is now called the theory of democratic elitism, best known to economists through the writings of Joseph Schumpeter. Its social function has been to facilitate the gradual but persistent isolation of the governing process from popular control, through the interposition of the expert between the electorate and the implementation of state action. The practice is well illustrated by the municipal reform movements of the progressive era. According to the historian Samuel Hays:

> Available evidence indicates that the source of support for reform in municipal government did not come from the lower or middle classes, but from the upper class. . . . The drama of reform lay in the competition for supremacy between two systems of decision making. One system based on ward representation . . . involved wide latitude for the expression of grass roots impulses . . . [In] the other . . . decisions arose from expert analysis and flowed from fewer and smaller centers outward to the rest of society. [pp. 152, 170]

More recently, economists have been in the vanguard of a similar process. In the ostensible interests of greater government rationality, advocates of program planning budgeting systems (*PPBS*) have sought (with mixed results) to limit the influence of the legislative branch upon the allocation of government resources. Benefit cost analysis and other public expenditure criteria serve in two quite distinct ways to separate the public from the resource allocation process. First, the determination of the value of a project is—as all of us who have practiced the art know—in the hands of the expert. Voters may be expected to have strongly felt, articulate, and informed opinions on the desirability of a new irrigation system or a new community hospital. On questions of the shadow price of labor or the social opportunity cost of capital, they are, not surprisingly, mute, leaving the actual public expenditure decision to the experts. Second, in practice public expenditure criteria tend to reintroduce in veiled form the very same market criteria which govern resource allocation in the private sector. And despite ideological assertions to the contrary, voting with dollars is profoundly undemocratic.

That economists and other "experts" have been an instrument in the erosion of democratic government may be accepted by some who nonetheless applaud the preoccupation of economists with the concerns of the powerful. The intimate connection between the state and the economics profession, it might be thought, allows economists to make a substantial contribution to solving our nation's pressing social problems. A rather fundamental question at issue here is: Can the goals of democracy and equality be achieved under capitalism through more enlightened and expert advice to policy makers? No, I argue, they cannot, and for two quite simple reasons. First, as many contemporary Marxists have argued, the inequalities and antidemocratic hierarchies which dominate our everyday lives are rooted in the capitalist system itself, not in preferences, technology, or ill-conceived state action. (See, e.g., Richard Edwards, Michael Reich, and Thomas E. Weisskopf.) They cannot be eradicated without overthrowing capitalism. Second, the fact that the state has not undermined capitalist institutions can hardly be chalked up to bad advice—from economists or any other "experts." The history of the state in the United States over the past century betrays its true role, one in which economists and other policy advisors are necessarily, if unwitting and even perhaps unwilling, accomplices. The state in capitalist society functions to reproduce the relations of production which define the capitalist order. (See Ralph Miliband, James Weinstein, and James O'Conner.) While state power is essential in any revolutionary process, at this stage of development of forces for radical change in the United States, the state cannot be used as an instrument for the destruction of capitalist institutions.

Yet, ironically enough, the continued extension of state power has in some essential respects strengthened the potential for radical change. In an earlier era, poverty, economic insecurity, alienation and other economic problems appeared to arise naturally from the shortcomings of the human species, or from the stinginess of nature: the laws of economics, to paraphrase Marx, appeared as immutable as the laws of nature. Under pressure from radical social movements—elements of organized labor, the urban poor, blacks—and with the ideological and technical guidance of liberals, the state has been forced to take responsibility for problems which it cannot solve without going beyond the confines of the capitalist system. The contradictions arising in the economic base of society have thus been in large measure displaced into the state sector. In its efforts to use the state to reproduce the capitalist order, the capitalist class has in true dialectical fashion lent strength to its own antithesis; it has revealed the political nature of the unsolved and unsolvable problems of advanced capitalism.

The resulting possibilities for socially concerned economists are immense. We can work with the various social movements which arise in the politicization of social problems: welfare recipients, tenants' organizations, third-world liberation movements, labor unions, food co-ops and other consumer groups, women's centers, student groups, prisoners' organizations, tax reform groups and the like. We can pursue our more conventional intellectual work with a dedication to demonstrate the unscientific nature of the capitalist ideology whose bankruptcy in practice has been demonstrated in the headlines of the past decade. This is no easy task, given the entrenched power of orthodoxy in the universities, but it is greatly aided by the evident failure in practice of the preconceptions of both the Priests and the Engineers. Moreover, as the Engineers in their pragmatic but largely fruitless search for social solutions depart ever farther from the conceptual home turf of the Priests the entire structure of capitalist economics loses its internal cohesion and becomes inconsistent in practice, incoherent to the "public" and to students, and increasingly vulnerable to attack. The weaknesses of conventional economics are thus but a reflection of the inability of capitalism to solve the problems it generates.

What then are the policy priorities for economists in the 1970's? To liberate ourselves from the view that beneficial social change can come through expert and enlightened advice to the powerful, and to take a position squarely with the vast majority of Americans in whose interest it is to overturn the structure of power. Our role as economists, I believe, is unavoidably political as well as scientific. Let us, then, be explicitly political, both as activists and intellectuals, working to realize the radical potential in the present social forces and ultimately to render state power an instrument in the hands of the people.

REFERENCES

R. C. Edwards, M. R. Reich, and T. E. Weisskopf, *The Capitalist System; a Radical Analysis of American Society,* Englewood Cliffs 1972.

H. Gintis, "Consumer Behavior and the Concept of Sovereignty: Explanations of Social Decay," *Amer. Econ. Rev.,* May 1972, *62,* 268–278.

S. Fine, *Laissez-Faire and the General-Welfare State,* Ann Arbor 1956.

S. P. Hays, "The Politics of Reform in Municipal Government in the Progressive Era," *Pacif. Northwest Quart.,* October 1961, 145–171.

E. Herman and R. DuBoff, "The New Economics: Handmaiden of Inspired Truth," *Rev. Rad. Polit. Econ.,* Aug. 1972, *4,* 54–84.

E. K. Hunt, "Economic Scholasticism and Capitalist Ideology," in E. K. Hunt and J. Schwartz, *A Critique of Economic Theory,* Baltimore 1972, 186–193.

R. Miliband, *The State in Capitalist Society,* New York 1969.

J. O'Connor, *The Fiscal Crisis of the State,* New York 1973.

J. Weinstein, *The Corporate Ideal and the Liberal State,* Boston 1968.

Some Secular Changes in Business Cycles

By GEOFFREY H. MOORE*

Although industrialized countries continue to have business cycles, such cycles have changed significantly in character. In what follows I shall describe some of these changes and point to their possible implications for research and policy.

Perhaps the most obvious change is that business recessions—periods of actual decline in economic activity—have become less frequent, shorter and milder. Interruptions to a steady rate of growth are more often simply slowdowns rather than actual declines in aggregate economic activity. This kind of shift can be observed in the business recessions identified by the National Bureau of Economic Research. On the whole, the five recessions of 1948–70 were shorter than the five recessions of 1920–38, produced smaller declines in output, income and employment, and were less widespread in impact. But recent recessions have been accompanied by higher rates of unemployment than might have been expected in view of other evidence attesting to their mildness.

One of the factors underlying this shift toward recessions of lesser severity, and one reason why it may be expected to persist, is the trend in the industrial composition of employment. Industries that normally experience larger percentage reductions in employment when recession hits are less important in the overall economic picture nowadays, while industries that often continue to expand right through recession have become more important.

Of the eleven major industrial sectors that account for total employment, seven experienced reductions averaging three percent or more during the five recessions of 1948–70 (Table 1). These seven sectors include manufacturing of durable goods like autos and appliances, with an average drop of 12 percent; mining, with an average drop of 10 percent; transportation and utilities, with an average drop of 5 percent; and farming, manufacturing of nondurable goods like textiles, construction, and federal employment, with drops of 3 to 4 percent. Employment in these seven sectors constituted more than half of total employment in 1955, but by 1972 their share had declined to about two-fifths. The other four major sectors—wholesale and retail trade; services; finance, insurance, and real estate; and state and local government—experienced much smaller declines or actual increases in employment during the five most recent recessions. They accounted for slightly less than half of total employment in 1955; by 1972 they accounted for three-fifths of the total.

In short, the industries that have contributed *most* to reduced employment during recession have shown little or no growth during the past fifteen years or so, while those that have contributed *least* to recession have grown much faster. The added stability has reduced the impact of recession upon total employment by something like one-third. If the 1955 distribution of employment among the eleven sectors had prevailed in all five recessions of

---

* Vice-President/Research, National Bureau of Economic Research, Inc., and Senior Research Fellow, Hoover Institution, Stanford University.

TABLE 1—ESTIMATED EFFECT OF EMPLOYMENT TRENDS TO 1985 ON CYCLICAL STABILITY OF EMPLOYMENT

	Annual Percent Change in Employment During Five Recessions, 1948–70	Percentage Distribution of Total Employment in			Annual Percentage Rate of Change in Employment	
		1955	1972	1985	1955–72	1972–85
Sectors with substantial percentage declines in employment during recessions						
Durable manufactures	−11.9	14.9	13.0	13.2	0.7	1.9
Mining	− 9.6	1.3	0.8	0.6	−1.5	−0.2
Transportation & utilities	− 4.5	6.6	5.5	5.0	0.5	1.0
Agriculture	− 3.8	9.8	4.0	1.8	−3.6	−4.5
Nondurable manufactures	− 3.5	11.5	9.6	8.7	0.5	1.0
Contract construction	− 3.1	5.4	5.1	4.8	1.2	1.4
Federal government	− 3.0	3.3	3.1	2.6	1.1	0.4
Total		52.8	41.1	36.7	0.1	0.9
Sectors with small percentage declines or with increases in employment during recessions						
Wholesale and retail trade	− 0.8	20.1	21.5	20.8	2.0	1.5
Services	+ 1.8	15.9	20.0	22.2	2.9	2.8
Finance, insurance and real estate	+ 2.4	4.0	5.0	5.5	2.9	2.5
State and local government	+ 4.5	7.2	12.4	14.9	4.9	3.2
Total		47.2	58.9	63.4	2.9	2.3
All Sectors	− 3.0	100.0	100.0	100.0	1.6	1.8
Estimated percent change during recession for all sectors,						
based on 1955 distribution of employment	− 2.7					
based on 1972 distribution of employment	− 1.7					
based on 1985 distribution of employment	− 1.4					

Source: Compiled by the National Bureau of Economic Research, Inc. from data published by the U.S. Bureau of Labor Statistics. The projection of employment to 1985 is given in Ronald E. Kutscher, p. 39.

The percent changes in employment during recessions are computed from three-month standings of seasonally adjusted data centered on business cycle peak and trough months, on the base of cycle averages running from trough to trough (except 1948–49, where the cycle base runs from peak to peak). The peak to trough periods are: Nov. 48–Oct. 49; July 53–Aug. 54; July 57–Apr. 58; May 60–Feb. 61; and Nov. 69–Nov. 79.

Note: Data used to compute recession changes are from the establishment survey (jobs) except for agriculture, where the household survey (persons) is used. The "all sector" series derived in this manner undergoes larger declines during recession than total civilian employment from the household survey, because of the greater sensitivity of the nonfarm establishment data than the household data to recession. The average declines during the five recessions, 1948–70, are:

Nonfarm, household survey (persons)	−1.1%
Nonfarm, establishment survey (jobs)	−2.9
Farm, household survey (persons)	−3.8
Total, household survey (persons)	−1.4
Total, nonfarm establishment and farm household surveys	−3.0

The data for the distribution of employment in 1955, 1972 and 1985 are based on the jobs concept, but differ from those used to measure recession changes because of the inclusion in the former of self-employed, unpaid family workers, and paid household employment, and because of other statistical and conceptual differences.

1948–70, the average reduction in total employment would have been 2.7 percent. With the 1972 distribution, the reduction would have been 1.7 percent.

If the projections of industrial employment to 1985 that have recently been published by the Bureau of Labor Statistics are anywhere near the mark, this source of employment stability will endure.[1] In fact, aggregate employment in the four recession-proof sectors (trade, services, finance and state and local government) is expected to grow, between 1972 and 1985, just about twice as fast as in the seven recession-prone sectors. The former will, by 1985, contribute nearly two-thirds of total employment; the latter, only about one-third. More people will be working in jobs that are relatively secure. If the distribution of employment among sectors in 1985 corresponds to the projection, an average recession will reduce total employment by 1.4 percent—approximately half the estimated 2.7 percent reduction based on the 1955 distribution.

---

[1] Kutscher, p. 39.

One of the implications of these employment trends is that future recessions are more likely to be in the nature of slowdowns in the rate of economic growth. This prospect has already led economists here and abroad to experiment with new methods—or to resurrect old methods—of identifying economic fluctuations. The interest in the *GNP* gap and other measures of capacity utilization, in rates of change in activity rather than levels, in trend-adjustment techniques, and in spectral analysis, stems in part from this shift. At the National Bureau, Ilse Mintz has identified growth cycles in the economies of West Germany and the United States, and other students have done similar work for Japan, Canada, Great Britain, and other countries. For the United States, Mintz's chronology includes three "growth recessions" in addition to the five already included in the National Bureau's business cycle chronology.[2] These occurred in 1951–52, 1962–63, and 1966–67. In each, real *GNP* growth dropped to an average annual rate in the range of 2.5 to 3.5 percent over periods of a year to a year and a half. Another such period of slow growth in real *GNP* began last spring, well before the energy crisis came upon us.

The policy implication that I particularly would like to draw from all this is that it has become more important than ever for policy makers to support basic research into the causes and consequences of these interruptions to steady growth. Indeed, this implication holds for anyone who is trying to foresee the outcome of current trends and the effects of efforts to move them in a more favorable direction. The emergence of a relatively new phenomenon is apt to create all sorts of uncertainties, some of which can be dispelled by research.

Let me mention a few points that are important enough or intriguing enough to call for further exploration. First, I've already noted that the trend towards milder recessions does not seem to show up in the level of the unemployment rate. Curiously, while changes in the rate have become smaller, the rate itself has become higher in both recession and prosperity. Is this due to the shift in the composition of the labor force towards greater participation by women and young persons who are prone to higher rates of unemployment? Has the existence and wider scope of unemployment benefits increased the incidence of unemployment? If the risk of losing a job has diminished because of the shift in the industrial composition of employment, why hasn't the unemployment rate come down commensurately?

Second, federal employment is not very stable over the cycle. Federal employment has declined in every one of the five recessions since 1948, and by percentages that rival those in the private sector (Table 2). During the intervening expansions, federal employment has increased. Although opinions differ as to whether in the long run federal employment should go up or down, scarcely anyone would agree that it should be pro cyclical. Why has it been pro cyclical? How can the timing of changes in federal employment be better controlled

---

[2] Mintz (1973).

TABLE 2—CHANGES IN PRIVATE AND PUBLIC EMPLOYMENT DURING RECESSION, 1948–70

Change, in thousands of persons
(Percentage changes in parentheses)

Recession Period		Private		State and
Peak	Trough	Total	Federal	Local
Nov. '48	Oct. '49	−2,514	−58	+168
		(−5.3)	(−2.7)	(+4.1)
Jul. '53	Aug. '54	−1,862	−120	+289
		(−3.9)	(−5.5)	(+6.9)
Jul. '57	Apr. '58	−2,675	−39	+176
		(−5.3)	(−1.8)	(+3.5)
May '60	Feb. '61	−1,054	−55	+177
		(−2.1)	(−2.5)	(+3.0)
Nov. '69	Nov. '70	−1,218	−65	+419
		(−2.2)	(−2.6)	(+5.2)

Source: See Table 1 for source and method of computation. See also Ebanks.

with a view to offsetting rather than augmenting employment fluctuations in the private economy?[3]

Third, wholesale prices seem to have become less responsive to changes in final demand. During each of the five recessions of the 1920's and 1930's, industrial wholesale prices declined by rates exceeding 5 percent per year. In contrast, during the five recessions of the past twenty-five years, industrial wholesale prices declined on the order of 5 percent in only one instance, 1948–49. In part, this is related to the fact that the postwar recessions were shorter and milder than prewar. But this does not seem to account for the entire shift in price behavior.[4] What other factors are involved? Will they persist or intensify? What are their implications for the control of inflation over the longer run?

Fourth, financial markets have become increasingly sensitive to economic fluctuations. Interest rates have moved in far wider swings in recent recession-recovery periods than formerly.[5] Stock prices, too, have become more sensitive not only to the standard business cycle recessions but also to the milder growth recessions. Indeed, stock price movements have exaggerated the relative severity of recent recessions. Thus the common impression that declines in the market forecast recession may prove unduly pessimistic. What is the relation, if any, between the shifts in interest rate behavior and stock price behavior? What factors underlie the greater sensitivity of stock prices and what implications do they have for this form of investment?

A fifth and final point is that growth recessions appear to be international in scope. It has long been known that the more severe business cycle contractions have been worldwide. The Great Depression of the early 1930's is the best known example, but the severe contractions of 1907, 1921, and 1938 are equally illustrative. But it is now clear that the mild growth recessions of 1951–52, 1962–63, and 1966–67 in the United States had counterparts in Europe. As a result, our exports, which used to fluctuate with demand conditions abroad and were more or less independent of domestic demand, now conform more closely to domestic conditions because foreign demand also is more nearly in tune with ours and are also now subject to wider cyclical fluctuation. As Mintz has noted, "The contrast between this shift and the dampening of the domestic business cycle is striking, and suggests that policies or structural changes conducive to international stabilization have been far less effective than those conducive to national stability." This remark, though written in 1967, seems especially apt today.

But the full implications of a more

---

[3] For an analysis of the cyclical behavior of federal as well as state and local government employment, see Walter Ebanks.

[4] Phillip Cagan has documented this change and explored some of its sources in a National Bureau study (1973).

[5] Cf. Cagan (1966, 1972).

closely integrated international business cycle, or growth cycle, remain to be explored and projected. Better current information in comparable form for different countries on the state of the business cycle is needed. A project recently launched by the National Bureau on International Economic Indicators will, we hope, help to fill this need. The subject has a bearing on international economic policy problems ranging over a very wide field—monetary relations, inflation, foreign trade, capital flows, migration, to say nothing of energy. All these topics seem likely to require a major share of policy-makers' attention during the 1970's, and they will need all the help they can get from economic research.

REFERENCES

P. Cagan, *Changes in the Cyclical Behavior of Interest Rates*, Occasional Paper 100, National Bureau of Economic Research, 1966.

———, "The Recent Cyclical Movements of Interest Rates in Historical Perspective," *Bus. Econ.*, Jan. 1972.

———, "Changes in the Recession Behavior of Wholesale Prices: The 1920's and Post-World War II," unpublished manuscript, Nov. 1973.

W. W. Ebanks, *Public Employment and Post World War II Economic Fluctuations: A Study of the Stabilizing Effects of Government Employment*, unpublished Ph.D. dissertation, New York University, 1973.

R. E. Kutscher, "The United States Economy in 1985: Projections of *GNP*, Income, Output and Employment, *Monthly Labor Review*, Dec. 1973, 39.

I. Mintz, *Cyclical Fluctuations in the Exports of the United States Since 1879*, National Bureau of Economic Research, New York 1967.

———, "Dating U.S. Growth Cycles," *Explorations in Economic Research*, forthcoming, 1973.

Policy Issues in the International Economy of the 1970's

By H. S. Houthakker*

There is room for argument whether the 1970's started on January 1, 1970, or January 1, 1971, but one thing is certain: neither date was significant for international economic policy. If we have to periodize by decades at all, then the 1970's probably began on August 15, 1971—and not only in international economics. On that date the United States in effect brought an end to the existing international monetary system by suspending the convertibility of the dollar into gold. The immediate purpose of this move was to bring about a worldwide realignment of par values and particularly a fall in the value of the dollar relative to other currencies. This purpose was accomplished in two or three stages, depending on how one interprets the appreciation of the European currencies during the summer of 1973. The results of that action are already exceeding the most optimistic predictions—our balance of payments, after causing us and other countries endless (though frequently exaggerated) concern during the 1960's, is now in solid surplus and likely to remain so for the remainder of the decade. The principal qualification that should be made relates to the energy problem, on which I shall say more later. Our surplus will not be burdensome as long as other countries can run down their previously accumulated dollar assets.

But the results of the August 15 decision have not had an equally satisfying impact on the international monetary system. We had to pull down the house because it was no longer adapted to the needs of the time, but putting something better in its place has not been easy. For nearly a year we have been living with floating rates, a device long favored by a majority of academic economists. So far floating rates have worked reasonably well; although there have been some excessive rate fluctuations of speculative origin, the growth of world trade and capital movements does not appear to have been impeded. However, we do not have *freely* floating rates, the system many economists advocate, but rather dirty or managed floating, with strong remnants of fixed rates, especially in Europe. Official intervention in the foreign exchange markets has not come to an end, nor is it likely to. The notion that the exchange rate can be left entirely to market forces is anathema to most governments who fear that this would result in a loss of control over their domestic policies. We still need a new monetary *system*, as distinct from the uncoordinated operations now being practiced by all major countries. Our government has made detailed and thoughtful proposals for a new system that would combine reasonable predictability of exchange rates with a capacity to make needed adjustments promptly and with a minimum of turmoil. These proposals have not been accepted by other countries, and since the United States can live with the present interim situation, we have not pressed them very hard. Perhaps we should take comfort in the French saying that only provisional arrangements can last, but the risks of disintegration of the world economy cannot be overlooked. Despite all its defects the Bretton Woods system had the merit of being universal, whereas the prevailing conditions encourage regional groupings such as the European "snake in the tunnel" experiment. This is the main reason why international monetary reform should remain on the agenda.

---

* Harvard University.

A related issue is international liquidity, particularly as it affects the worldwide inflation we are now going through. In the 1960's there was considerable agitation over an alleged shortage of liquidity, in response to which the Special Drawing Rights were created. Actually the evidence for such a shortage was never impressive; while net official reserve assets may not have grown much, there was a huge creation of private international liquidity in the Eurodollar market. Moreover, official liquidity would have been more than adequate if greater reliance had been placed on exchange rates in the adjustment of imbalances. Now that exchange rates are being used more freely, the need for official liquidity is much reduced, and for the time being we should abstain from creating more. In the meantime, however, private international liquidity continues to expand. Since much of it is footloose, it does not show up in published money supply data, which in any case show excessive growth. Nevertheless it almost certainly contributes powerfully to the inflationary pressures that no nation has succeeded in keeping under control. If the world is ever going to take hold of the general price level, the question of international liquidity needs a more searching examination than it has had in the past.

Apart from the problem of inflation, which is only partly of an international nature, the most serious international problems are likely to be in particular markets. To no one's surprise I shall first say something about the petroleum market. The basic fact, as every economist should know, is not that the world is running out of energy resources, as the Club of Rome would have us believe. We are not running out, and even if we were, higher prices could easily take care of that since both supply and demand are price-elastic. The problem, rather, is the attempt by the oil-producing countries to cartelize the supply side. This attempt is now coming to a climax, and it would be rash indeed to make predictions about its ultimate success. The embargo by certain Arab producers against the United States and the Netherlands was proclaimed on political grounds, but it is already clear that its main effects will be economic. The world petroleum market is fairly well integrated; if one or two countries are denied access to certain sources of supply, they will simply outbid the other importers for alternative supplies. Those importers who continue to receive Arab oil, therefore, will get less from other sources, and the net result is a more or less equal reduction of supply to all countries, whether friendly to the Arabs or not. This reallocation is currently being carried out by the international oil companies since there is no open market in crude oil, but these companies can hardly be blamed for doing what elementary economics dictates. Even though the favored importers are therefore receiving less oil (assuming the embargo is effective and the cutback is not fully made up by increased supplies from other producers), they may still have an advantage over the "hostile" importers if they get the Arab oil at a lower price. However, recent pricing actions by the Organization of Petroleum Exporting Countries (*OPEC*) would seem to have wiped out this advantage.

Indeed, it now appears that the selective embargo, far from being inspired by lofty motives of Arab solidarity, was essentially a move to create discord among the importing countries, thus torpedoing any contemplated consumer cartel and making it easier for *OPEC* to double and triple the price. In this respect it has so far been successful. Although members of the Organization for Economic Cooperation and Development (*OECD*) apparently had a prior agreement to share oil in case of a supply interruption, our allies across the Atlantic and the Pacific broke ranks at the first shot. The European Community also turned out to be a sham when one of its members was cut off from Arab oil. Now our allies are being presented by their new friends with the bill for their shortsightedness; the producers are gradually terminating the embargo which has served its real economic purpose without accomplishing any of its stated political goals. The producers' main concern is now what to do with their vastly increased revenues.

Of course this is not the end of the story. There may be markets where the price can be doubled or tripled without changing the balance of supply and demand, but the petroleum market is certainly not one of them. The incentives for increasing production are tremendous, and no matter what the Club of Rome may say, the potential for greater output exists, not only in the currently exporting countries, but also in the United States. In the very short run not much can be done, but at prevailing *OPEC* prices we are likely to see a surplus of oil within the next year or two. Normal responses to higher prices on the demand side will work in the same direction. The producer cartel will then have to adopt prorationing if it wants to keep the price up. The abstention of Iraq and Libya from the present embargo suggests that cohesion among the exporters may be as hard to achieve as cohesion among the importers.

This episode has wider ramifications. The United States has made self-sufficiency in energy an official policy goal, though one must hope that this means a *potential* for self-sufficiency to forestall extortion by *OPEC*, rather than actual autarchy which would cost too much and disrupt our international economic relations. The balance of payments of Europe, Japan and most of the developing countries will be strained by higher oil prices, and this may in turn affect our exports to these countries; at the same time, however, it will make their exports less competitive in the *U.S.* market, so that the net effect on our balance of payments is likely to be small.

I have not said anything about agriculture or the more traditional trade problems, particularly tariffs and nontariff barriers to trade. The message is the same as with respect to the problems I have already discussed: without more effective international cooperation the world economy may break into regional blocks. The atmosphere has deteriorated and meaningful new agreements among the Western countries appear to be harder and harder to negotiate. The Europeans are increasingly looking inward to the construction of their Community, but the results achieved have not been commensurate with the negotiating efforts. Japan has not yet found a posture befitting its importance in the world economy. Our own strong-arm tactics over the dollar—even though justified by their results and by the tendency of other countries to turn a blind eye to our problems—have hardly promoted international understanding; our autarchic pronouncements about energy, equally understandable in the circumstances, are also open to misinterpretation. To reverse these tendencies will be the major challenge to the international economy in the 1970's.

Some Reasons for the Maximin Criterion

By JOHN RAWLS*

Recently the maximin criterion of distributive equity has received some attention from economists in connection with the problem of optimal income taxation.[1] Unhappily I am unable to examine the merits of the criterion from the standpoint of economic theory, although whether the criterion is a reasonable distributive standard depends importantly on the sort of examination that only economists can undertake.

What I shall do is to summarize briefly some of the reasons for taking the maximin criterion seriously. I should emphasize that the maximin equity criterion and the so-called maximin rule for choice under uncertainty are two very different things. I shall formulate the reasons for the equity criterion so that they are completely independent from this rule.

In *A Theory of Justice* I have considered the maximin criterion as part of a social contract theory. Here I must assume a certain familiarity with this conception.[2] One feature of it might be put this way: injustice exists because basic agreements are made too late (Richard Zeckhauser). People already know their social positions and relative strength in bargaining, their abilities and preferences, and these contingencies and knowledge of them cumulatively distort the social system. In an attempt to remedy this difficulty, contract theory introduces the notion of the original position. The most reasonable principles of justice are defined as those that would be unanimously agreed to in an appropriate initial situation that is fair between individuals conceived as free and equal moral persons.

In order to define the original position as fair in this sense, we imagine that everyone is deprived of certain morally irrelevant information.[3] They do not know their place in society, their class position or social status, their place in the distribution of natural assets and abilities, their deeper aims and interests, or their particular psychological makeup. Excluding this information insures that no one is advantaged or disadvantaged in the choice of principles by natural chance or social contingencies. Since all are in this sense similarly situated and no one knows how to frame principles that favor his particular condition, each will reason in the same way. Any agreement reached is unanimous and there is no need to vote.

Thus the subject of the original agreement is a conception of social justice. Also, this conception is understood to apply to the basic structure of society: that is, to its major institutions—the political constitution and the principal economic and social arrangements—and how they fit together into one system. The application of the maximin criterion to optimal income taxation is, then, perfectly in order, since an income tax is part of the basic structure. But the maximin criterion is not meant to apply to small-scale situations, say, to how a doctor should treat his patients or a university its students.[4] For these situations different principles will presumably be necessary. Maximin is a macro not a micro principle. I should add that the criterion is unsuitable for determining the just rate of savings; it is intended to hold only within generations (Rawls, sec. 44, pp. 291–92, Kenneth J. Arrow, 1973a, and Robert M. Solow).

---

* Professor of philosophy, Harvard University. I am grateful to Robert Cooter and Richard Zeckhauser for their very instructive comments which enabled me to improve these remarks and saved me from several mistakes.

[1] See A. B. Atkinson, Martin Feldstein, Yoshitaka Itsumi, and Edmund S. Phelps.

[2] In economics this sort of theory was reintroduced by James Buchanan and Gordon Tullock.

[3] A similar idea is found in J. C. Harsanyi (1953).

But what alternative conceptions are available in the original position? We must resort to great simplifications in order to get our bearings. We cannot consider the general case where the parties are to choose among all possible conceptions of justice; it is too difficult to specify this class of alternatives. Therefore we imagine that the parties are given a short list of conceptions between which they are to decide.

Here I can discuss only two pair-wise comparisons. These are designed to reflect the traditional aim of contract theory, namely, to provide an account of justice that is both superior to utilitarianism and a more adequate basis for a democratic society. Therefore the first choice is between a conception defined by the principle that average utility is to be maximized and a conception defined by two principles that express a democratic idea of justice. These principles read as follows:

1. Each person has an equal right to the most extensive scheme of equal basic liberties compatible with a similar scheme of liberties for all.
2. Social and economic inequalities are to meet two conditions: they must be (a) to the greatest expected benefit of the least advantaged members of society (the maximin equity criterion) and (b) attached to offices and positions open to all under conditions of fair equality of opportunity.

I assume that the first of these takes priority over the second, but this and other matters must be left aside. (For more on this, see Rawls, pp. 40–45 and 62–65.) For simplicity I also assume that a person's utility is affected predominantly by liberties and opportunities, income and wealth, and their distribution. I suppose further that everyone has normal physical needs so that the problem of special health care does not arise.[5]

Now which of these conceptions would be chosen depends on how the persons in the original position are conceived. Contract theory stipulates that they regard themselves as having certain fundamental interests, the claims of which they must protect, if this is possible. It is in the name of these interests that they have a right to equal respect and consideration in the design of society. The religious interest is a familiar historical example; the interest in the integrity of the person is another. In the original position the parties do not know what particular form these interests take. But they do assume that they have such interests and also that the basic liberties necessary for their protection (for example, freedom of thought and liberty of conscience, freedom of the person, and political liberty) are guaranteed by the first principle of justice.

Given these stipulations, the two principles of justice would be chosen For while the principle of utility may sometimes lead to a social order securing these liberties, there is no reason why it will do so in general. And even if the principle often does, it would be pointless to run the risk of encountering circumstances when it does not. Put formally, each must suppose that the marginal utility of these fundamental interests is infinite; this requires anyone in the original position to give them priority and to adopt the two principles of justice.

---

[4] This affects the force of Harsanyi's counterexamples (Harsanyi, 1973).

[5] Part of the justification for these assumptions is given by the notion of primary goods. See Rawls, sec. 15. Of course, there are still difficulties (Arrow, 1973b, p. 254).

This conclusion is strengthened when one adds that the parties regard themselves as having a higher-order interest in how their other interests, even fundamental ones, are regulated and shaped by social institutions. They think of themselves as beings who can choose and revise their final ends and who must preserve their liberty in these matters. A free person is not only one who has final ends which he is free to pursue or to reject, but also one whose original allegiance and continued devotion to these ends are formed under conditions that are free. Since the two principles secure these conditions, they must be chosen.

The second pair-wise comparison is far more difficult. In this case the choice is between the two principles of justice and a variant of these principles in which the utility principle has a subordinate place. To define this variant, replace the second principle by the following: social and economic inequalities are to be adjusted so as to maximize average utility consistent with fair equality of opportunity. The choice between this variant and the two principles is more delicate because the arguments from liberty can no longer be made, at least not so directly. The first principle belongs to both conceptions, and so the operation of the utility principle is hedged by basic rights as well as fair equality of opportunity.

One reason favoring the two principles of justice is this. From the standpoint of the original position, the parties will surely be very considerably risk-averse; if we ask how risk-averse, we might say not less than that of most any normal person. Of course, this is extremely vague; but if we assume that utility is estimated from the standpoint of individuals in society and represents, as the classical utilitarians believed, a quantity ascertainable independent of choices involving risk, then, given the crucial nature of the decision in the original position, the claims of the utility principle seem quite dubious. On the other hand, if we suppose that utility is measured from the original position and takes account of risk, the utility criterion may not differ much from maximin. The standard of utility approaches maximin as risk aversion increases without limit (Arrow, pp. 256–57). So, either way, the original position pushes us toward maximin. However, in weighing the second pair-wise comparison, I assume that, based on considerations of risk aversion alone, there is a significant difference between the two conceptions. Thus the problem is to identify other attractive features of the maximin criterion that tip the balance of reasons in its favor.

First, much less information is needed to apply the maximin criterion. Once the least-favored group is identified, it may be relatively easy to determine which policies are to their advantage. By comparison it is much more difficult to know what maximizes average utility. We require a fairly precise way of comparing the utilities of different social groups by some meaningful standard, as well as a method of estimating the overall balance of gains and losses. In application this principle leaves so much to judgment that some may reasonably claim that the gains of one group outweigh the losses of another, while others may equally reasonably deny it. This situation gives those favored by existing inequalities an opportunity to exploit their advantage so that, as a result, inequalities are likely to be excessive, undermining the justice of the system.

A further consideration is this: a distributive criterion is to serve as a public principle. Citizens generally should be able to understand it and have some confidence that it is realized. Pattern criteria, those that require the actual distribution to exhibit certain ascertainable features, do well by the test of publicity. Of these, strict equality (equal division) is the sharpest principle. The trouble with pattern criteria is that sharpness is not the only desideratum, and they often have little else to commend them. On the other hand, the utility principle is not sharp enough: even if it were satisfied, there could be little public confidence that this is indeed the case. The maximin criterion has sufficient sharpness; at the same time it is efficient while strict equality is not.

Another ground supporting the maximin criterion is based on the strains of commitment: in the original position the parties are to favor those principles compliance with which should prove more tolerable, whatever their situation in society turns out to be. The notion of a contract implies that one cannot enter into an agreement that one will be unable to keep. By this test, also, maximin seems superior, for the principles chosen would regulate social and economic inequalities in the basic structure of society that affect people's life-prospects. These are peculiarly deep and pervasive inequalities and often hard to accept.

Looking first at the situation of the less advantaged, the utility principle asks them to view the greater advantages of others who have more as a sufficient reason for having still lower prospects of life than otherwise they could be allowed. This is an extreme demand psychologically; by contrast, the maximin criterion assures the less favored that inequalities work to their advantage. The problem with maximin would appear to lie with those who are better situated. They must accept less than what they would receive with the utility principle, but two things greatly lessen their strains of commitment: they are, after all, more fortunate and enjoy the benefits of that fact; and insofar as they value their situation relatively in comparison with others, they give up that much less. In fact, our tendency to evaluate our circumstances in relation to the circumstances of others suggests that society should be arranged so that if possible all its members can with reason be happy with their situation. The maximin criterion achieves this better than the principle of utility.

I have noted several reasons that support the maximin criterion: very considerable normal risk-aversion (given the special features of the original position), less demanding information requirements, greater suitability as a public principle, and weaker strains of commitment. Yet no one of them is clearly decisive by itself. Thus the question arises whether there is any consideration that is compelling. I want to suggest that the aspirations of free and equal personality point directly to the maximin criterion.

Since the principles of equal liberty and fair equality of opportunity are common to both alternatives in the second comparison, some form of democracy obtains when either alternative is realized. Citizens are to view themselves as free and equal persons; social institutions should be willingly complied with and recognized as just. Presumably, however, certain social and economic inequalities exist, and individuals' life-prospects are significantly affected by their family and class origins, by their natural endowments, and by chance contingencies over the course of their lives. We must ask: In the light of what principle can free and equal moral persons permit their relations to be affected by social fortune and the natural lottery? Since no one deserves his place in the distribution of talents, nor his starting place in society, desert is not an answer. Yet free and equal persons want the effects of chance to be regulated by some principle, if a reasonable principle exists.

Now when the maximin criterion is followed, the natural distribution of abilities is viewed in some respects as a collective asset. While an equal distribution might seem more in keeping with the equality of free moral persons, at least if the distribution were a matter of choice, this is not a reason for eliminating natural variations, much less for destroying unusual talents. To the contrary, natural variations are recognized as an opportunity, particularly since they are often complementary and form a basis for social ties. Institutions are allowed to exploit the full range of abilities provided the resulting inequalities are no greater than necessary to produce corresponding advantages for the less fortunate. The same constraint holds for the inequalities between social classes. Thus at first sight the distribution of natural assets and unequal life-expectations threatens the relations between free and equal moral persons. But provided the maximin criterion is satisfied, these relations may be preserved: inequalities are to everyone's advantage and those able to gain from their good fortune do so in ways agreeable to those less favored. Meeting this burden of proof reflects the value of equality.

Now the maximin criterion would conform to the precept "from each according to his abilities, to each according to his needs" if society were to impose a head tax on natural assets. In this way income inequalities could be greatly reduced if not eliminated. Of course, there are enormous practical difficulties in such a scheme; ability may be impossible to measure and individuals would have every incentive to conceal their talents. But another difficulty is the interference with liberty; greater natural talents are not a collective asset in the sense that society should compel those who have them to put them to work for the less favored. This would be a drastic infringement upon freedom. But society can say that the better endowed may improve their situations only on terms that help others. In this way inequalities are permitted in ways consistent with everyone's self-respect.

I have attempted a brief survey of the grounds for the maximin criterion. I have done this because historically it has attracted little attention, and yet it is a natural focal point between strict equality and the principle of average utility. It turns out to have a number of attractive features. But I do not wish to overemphasize this criterion: a deeper investigation covering more pair-wise comparisons may show that some other conception of justice is more reasonable. In any case, the idea that economists may find most useful in contract theory is that of the original position. This perspective can be defined in various ways and with different degrees of abstraction and some of these may prove illuminating for economic theory.[6]

REFERENCES

K. J. Arrow, "Rawls's Principle of Just Saving," Instit. for Math. Stud. in the Soc. Sciences, Stanford, tech. rep. no. 106, Sept. 1973a.

———, "Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice," *J. Phil.*, May 1973b, *70*, 254.

A. B. Atkinson, "How Progressive Should Income Tax Be?" in J. M. Parkin, ed., *Essays in Modern Economics*, London 1973.

J. Buchanan and G. Tullock, *The Calculus of Consent*, Ann Arbor 1962.

---

[6] See Zeckhauser for an illustration of the use of this sort of framework.

M. **Feldstein**, "On the Optimal Progressivity of the Income Tax," Harvard Instit. of Econ. Res., disc. pap. 309, July 1973.

J. C. **Harsanyi**, "Cardinal Utility in Welfare Economics and in the Theory of Risk Taking," *J. Polit. Econ.*, Oct. 1953, *61*, 434–35.

———, "Can the Maximin Principle Serve as a Basis for Morality?" Center for Res. in Manage. Sc., Berkeley, work. pap. no. CP-351, May 1973, 8–9.

Y. **Itsumi**, "Distributional Effects of Income Tax Schedules," *Rev. Econ. Stud.*, forthcoming.

E. S. **Phelps**, "The Taxation of Wage Income for Economic Justice," *Quart. J. Econ.*, Aug. 1973, *87*, 331–54.

J. **Rawls**, *A Theory of Justice*, Cambridge, Mass. 1971.

R. M. **Solow**, "Intergenerational Equity and Exhaustible Resources," dept. of econ., work. pap. no. 103, Cambridge, Mass. Feb. 1973.

R. **Zeckhauser**, "Risk Spreading and Distribution," in H. M. Hochman and G. E. Peterson, eds., *Political Economics of Income Distribution*, New York 1974.

Achieving the Just Polity

By Dennis C. Mueller*

Any discussion of the distribution of income immediately raises issues of equity and justice. Most theories of justice are built around the prescription that one put one's self in the other fellow's shoes—more to the point, that the rich man place himself in the position of the poor.

In this paper, I consider two models of justice and ask about the *feasibility* of such mental experiments, i.e., whether individuals will in fact make them. I assume that the desirability of such experiments is universally accepted.

Consider first a world of perfect information and mobility. Clubs (polities) form to provide public goods for their members. Up to a point, an additional membership raises the average net benefits to all club members due to the joint supply aspects of all public goods. Beyond this point, either these aspects are totally exhausted or the benefits from them are offset by congestion and policing problems. (If no such point is ever reached, there is but one, all-inclusive club.) The existing membership of any club is free to define its rules so as to attract or exclude other members. Individuals are free to quit and join another club, and they have identical tastes and abilities. Clubs are formed under conditions of free entry and exit. The total population is large relative to the size of any club. Under these conditions, the optimal set of rules defining membership provides that each member receives the identical (average) benefits. Until optimum club size is attained, the current members have, if anything, an incentive to discriminate in *favor* of new members, since their addition raises the average benefits to existing members. (The assumption of a large population relative to club size ensures that no club's growth is arrested in this position). At the optimum, it does not pay current members to discriminate in favor of new members, the marginal and average benefits of club membership being equal. Any attempt to offer new members less than the average benefits from club memberships, or to form coalitions which discriminate against other members, is thwarted by their ability to exit and form other clubs to obtain the same average benefits. Thus, in a world of perfectly competitive clubs, fair rules for the sharing of total benefits (which in the case of identical tastes and abilities would mean that each member receives the same average benefits) would emerge as an inevitable byproduct of the competition among clubs for membership, as each tried to obtain or maintain the optimal club size. If tastes and abilities are not identical, each member receives benefits tailored to his own attributes, and competition among clubs ensures a form of horizontal equity across clubs.

This analysis is essentially an alternative interpretation of Charles M. Tiebout's model. Tiebout emphasized the achievement of allocative efficiency: individuals preferring low expenditure on schools gather in one community, those preferring high levels in another. I stress the achievement of equity: all individuals with identical incomes and tastes for education must be treated equally. To attract and retain club members, each current member is forced by competition to place himself in the positions of potential and other existing club members and thereby to establish benefits at least as good as those obtainable from other clubs. Thus, in a decentralized world of competitive clubs the invisible hand achieves the golden rule.

---

* Associate professor of economics, Cornell University. This paper was written while I was a postdoctoral fellow at the Center for Public Choice at Virginia Polytechnic Institute and State University. Stimulation for many of its ideas was obtained from the Center's members. Particular thanks are due to Thomas Borcherding James Buchanan, Mark R. Killingsworth, James Meade, Robert Tollison, and Gordon Tullock.

But the Tiebout model's assumptions of perfect competition, mobility, and information are unrealistic. This criticism is particularly valid if the model is to be used to consider whether *competitive justice* will be provided at, say, the nation-state level. People cannot easily envisage what citizenship in Germany, Switzerland, Russia, Peru, or China would be like, and they cannot immediately become citizens of these countries. Nor can one describe these countries as competing with equal vigor for new citizens.

In the second model I wish to consider, the members of the polity are envisaged as adopting a social contract or constitution while being, because of the long-run nature of these contracts, uncertain as to their own position within them. Because of this uncertainty each individual is forced through rational self-interest to envisage the impact of any rule on all other citizens. In deciding on distributional norms the rich must envisage the impact of such norms on the poor because they or their descendants may someday be poor. This notion of justice as fairness has been discussed by James Buchanan and Gordon Tullock, John Harsanyi, Abba Lerner, Dennis Mueller (1973), and most exhaustively and prominently by John Rawls. The model lies at the other pole from Tiebout's for it assumes (usually implicitly) that individuals are totally immobile and (explicitly) that they have no "morally irrelevant" information about, e.g., their own and others' preferences and current and likely future positions in the distribution of income. Individuals are immobile in that each knows or assumes he will be a member of the polity and is thus motivated out of self-interest to adopt a set of rules that maximize his expected welfare. On the other hand, he must weigh the impact of these rules on *all* members of the community because he is uncertain as to the position in that community in which he is or will be.

While the Tiebout model is unrealistic in assuming too much mobility and information, the justice-as-fairness model seems unrealistic in assuming too little. Rawls is the most extreme in this regard, denying individuals knowledge of even the generation to which they belong, but all the authors mentioned above assume that individuals have no "morally irrelevant" information, e.g., that a rich man will not know that he is (or has a greater chance of remaining) rich rather than (becoming) poor, that a white man will not know that he is (and will remain) white rather than black, etc.

Both models, however, do have a number of common elements. In both, justice, if it appears, does so as a means of achieving some other end, not as an end in itself. The *raison d'être* of the state is the protection of the lives and property of its citizens and the provision of various other public goods. Although one can envisage individuals forming a club to insure against changes in relative positions of income and wealth, such a club would logically have to be formed *after* the creation of a state which established property rights and thereby created something to insure (Buchanan, 1973). Thus the social contract or constitution would not be written with the original intention of establishing justice. Nevertheless, it might incorporate impartial rules in it as a means for maximizing the expected utility of the constituents under uncertainty or under conditions of free exit and entry.[1]

Since in either case the polity is formed to provide a certain set of public goods, not justice, under what circumstances is a just polity feasible? The boundaries which are optimal for achieving allocational efficiency in the provision of public goods may not be the same as the boundaries for achieving just rules. The former are determined by the technological characteristics of the public goods and the distribution of the population (Mueller, 1971), the latter by the knowledge, mobility, and distribution characteristics of the population. Thus, that we consider the United States a single community when it comes to providing national defense does not suggest that we *will* (as against *ought to*) consider the United States a single community for undertaking redistribution.

One can, of course, envisage situations in which the collective provision of a public good produces the knowledge and understanding necessary for a just society. Consider a volunteer fire department. Here the actual physical and social contact with one's neighbors may create or reinforce feelings of empathy. The uniform uncertainty over whose house may catch on fire next guarantees that each will do "his fair share" with respect to not only fire fighting but possibly other community activities as well.

It is far less likely that the joint consumption of the services of the modern fire department establishes any common social bonds, however. Many of the modern public goods are consumed with little or no physical contact, e.g., the benefits of environmental protection, travel on highways, and national defense.

In general, both the geographic and issue dimensions of the polity for allocative efficiency seem to be increasing. Technological changes more often than not seem to increase the optimal scale of production, e.g., the replacement of fossil fuels by atomic energy in electricity production. Urbanization increases the dimensions of government action by weakening the effectiveness of informal sanctions in curbing antisocial behavior. The citizen of a small town is restrained from littering the town square by the knowledge that his behavior encourages others to litter and may incur the disapprobation of his neighbors. In Central Park the number of visitors is so large that one can litter in anonymity and with a clear conscience, knowing that one's contribution has no appreciable impact on the total amount of litter in the park. In a large city, such antisocial behavior can be curbed only if the watchful eyes of the small town are replaced by those of the policeman (Buchanan, 1965). The increasing mobility of individuals, goods, and factories extends the boundaries of interaction and exchange and thereby the dimensions of necessary collective action.

Increasing mobility by itself, of course, increases the potential for achieving a Tieboutian justice. Unfortunately, perhaps, mobility alone cannot be relied upon to produce a just society. First, even in a Tieboutian world, some knowledge of the tastes and abilities of others is needed to entice them into the polity, and hence the problem of acquiring this knowledge remains. Second, Tieboutian justice *is* a concept which is specific to an individual's tastes and abilities. It ensures horizontal equity but not vertical equity. If polities are unwilling to compete for the aged, the blind, or the anarchists, then these groups may find themselves still receiving a very small slice of the pie even under full mobility. Third, and most importantly, mobility itself creates its own externalities, necessitating intervention by some central authority and thereby eliminating the possibility of a fully decentralized resolution of all social issues à la Tiebout (Buchanan and Charles J. Goetz). In a highly mobile society we are all members (or potential members) of the same community on some issues.

---

[1] Some private clubs or charities may be formed under some form of insurance motive (Geoffrey Brennan). They also may be formed to obtain a Hochman-Rogers-type redistribution. In the latter case, the goal is still not the creation of justice, but simple utility maximization where the club's goal (public good) is a redistribution away from or toward some other group whose utility is in the club member's preference function. As Harold Hochman and James Rogers correctly argue, this type of redistribution is best treated as a Pareto reallocation of resources.

Thus the full achievement of a just society probably does require the adoption of some form of Rawlsian justice in which each individual assumes he might actually change positions with other individuals. The capacity to make such a mental experiment depends on one's imagination and knowledge of those with whom he might exchange places. For example, suppose I happen across a man beating a dog. Is this an immoral act? Under the Rawlsian code of ethics, one cannot say. Right and wrong are determined by placing oneself in the position of others under situations of uncertainty. If it is impossible for a man to undergo the mental experiment of envisaging being a dog, and Rawls explicitly refuses to assume that he can (pp. 504–12), then one cannot condemn him for treating a dog differently than a man. If he is deficient, the deficiency lies in his powers of imagination, not his moral fiber.

Can the same argument be extended to one man's treatment of another? Is it possible that one person is so devoid of information about another that he simply cannot mentally assume the position of the other and pretend there is any uncertainty over which position he will occupy? If this is true, is it more meaningful or in any way helpful to label him immoral rather than unimaginative? Is not the remedy to the situation much the same: i.e., provide the first man with more knowledge of and contact with the second, and also increase the probability that he might actually be in the second's position?

Traditionally the bulk of knowledge of others has probably come through direct personal contact. Thus most value systems favor those with whom one lives and works. Sociopolitical systems are often built around either some blood relationship (family, tribe, ethnic subgroup, race) or geographic proximity (household, neighborhood, city, region, nation), or both. Social mobility (and therefore uncertainty about future position) is usually greater within a cultural or geographic grouping rather than across these boundaries.

Knowledge of others, like any other human capital, can be acquired at either the intensive or extensive margins: one can either obtain additional knowledge of the same people or gather information, presumably in ever smaller quantities per person, about more and more people. Depending on which method of gathering information were employed, one would expect quite different value systems and codes of justice. Where a lot of information is acquired about a small group of people, one anticipates intense loyalty and justice exhibited toward those within the group, but perhaps different codes applicable to outsiders. This is the behavior one associates with social orderings built around the family, tribe, or town. One might view an injury against a member of one's family or tribe as an injury against oneself, but not react at all when people outside the close-knit social group experience extreme hardship. The possibility that one might exchange positions with another person within a close-knit group is obviously much greater than with outsiders, and this reinforces the bonds of empathy. Indeed it is probably a necessary condition for their creation. In a tribal society, one tribesman may defend another from outside attack precisely because he knows that he may be under attack next and wishes his tribesman to come to his support. In a rigid caste or slave society, in contrast, one may have much contact among people who have little or no chance of exchanging positions. Members of different castes may therefore have little empathy or understanding for one another. Not surprisingly, one typically finds different codes of justice applying across caste lines than within them. To maintain these lines and the low possibility of exchanging positions, one would naturally expect to find strict taboos against "mingling" and intermarriage.

Alternatively one can acquire knowledge of large numbers of people by brief contact through travel and by living in an urban environment. This knowledge can be extended by study and reflection. One can read about other people, reflect upon and generalize from one's own limited observations and thereby perhaps expand one's knowledge of others at both the extensive and intensive margins. Mobility, social intercourse, and education are thus the traditional liberal means for achieving a better (more equitable) society (L. J. Fein).

Toward the end of his book, Rawls paints an optimistic picture, reminiscent of Kant, of a society evolving into ever more just states. With increasing social mobility and knowledge about others, more and more people undertake the mental experiment of exchanging positions, and each person finds himself being treated more justly. This leads him to treat others more justly, and so on. The dimensions of the just society grow until it encompasses all people in all generations. For the just society Rawls describes to become reality, it would seem that knowledge of others must be acquired at both the intensive and extensive margins, however, and probably at increasing rates. The strong bonds of identification that are established in a close-knit family must be extended to include, eventually, all of mankind. Unfortunately this seems unlikely to happen. The acquisition of knowledge on the extensive margin must come at the expense of knowledge on the intensive margin. If we are to learn more of those far away, we must inevitably neglect those close at hand. Thus, our empathy for others increases in breadth while it reduces in depth. We become a nation paying ever larger percentages of national income into social security, while we individually abandon our parents and grandparents.

Some modern developments do enhance the individual's ability mentally to change places with others. The mass media is an obvious example. Quickening technological change and geographic and social mobility tend to increase the range of uncertainty over possible changes in position. An increasing percentage of the population can now easily envisage living in California or Maine, or even Europe and Japan; being unemployed; having a college degree or not having one; and so on. Urbanization replaces the sometimes biased dictates of peer group pressure in determining mores with the impartiality of codified justice administered by the courts and bureaucracies. To the extent that the city is a pot, melting away cultural differences, urbanization reduces the time required to understand others. Specialization increases one's sense of interdependence and perhaps, as Emile Durkheim argued, solidarity with the larger community. The extension of the market economy through the growth of international trade, with its concomitant destruction of economic systems built on caste and tradition, brings about a more egalitarian and impartial view of man (Buchanan and Tullock, p. 80). The spread of Western technology and with it Western culture and values reduces the costs of knowing the preferences of others. If, in adopting our technologies and buying our products, the rest of the world does inevitably adopt our culture, the problem of understanding those from a different background is tautologically removed. It may well be that once one has met one post-industrial man, one has met them all.

This possibility raises a host of additional questions about the characteristics and feasibility of the just society—and at least one disturbing paradox. The developments—technological change, specialization, mobility, increasing education, urbanization—that produce the universal values that must underlie a just society are the same developments sociologists and psychologists list as causes of the identity crisis and anomie that supposedly characterize man in the modern society. One is left wondering whether the just society toward which we are heading is one of empathy and impartiality, or loneliness and indifference. The line between them sometimes seems very fine.

REFERENCES

G. Brennan, "Pareto Desirable Redistribution: The Non-Altruistic Dimension," *Pub. Ch.*, Spring 1973, *14*, 43–67.

J. M. Buchanan, *The Limits of Liberty: Between Anarchy and Leviathan*, Blacksburg 1973.

———, "Ethical Rules, Expected Values, and Large Numbers," *Ethics*, Oct. 1965, *76*, 1–13.

——— **and C. J. Goetz,** "Efficiency Limits of Fiscal Mobility: An Assessment of the Tiebout Model," *J. of Pub. Econ.*, 1972, *1*, 25–43.

——— **and G. Tullock,** *The Calculus of Consent*, Ann Arbor 1962.

E. Durkheim, *The Division of Labor in Society*, Toronto 1933.

L. J. Fein, *The Ecology of Public Schools: An Inquiry into Community Control*, New York 1971.

J. C. Harsanyi, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *J. of Polit. Econ.*, Aug. 1955, *43*, 309–26.

H. M. Hochman and J. D. Rogers, "Pareto Optimal Redistribution," *Amer. Econ. Rev.*, Sept. 1969, *59*, 542–57.

A. P. Lerner, *The Economics of Control*, New York 1944.

D. C. Mueller, "Fiscal Federalism in a Constitutional Democracy," *Pub. Pol.*, Fall 1971, *19*, 567–93.

———, "Constitutional Democracy and Social Welfare," *Quart. J. of Econ.*, Feb. 1973, *87*, 60–80.

J. Rawls, *A Theory of Justice*, Cambridge, Mass. 1971.

C. M. Tiebout, "A Pure Theory of Local Government Expenditures," *J. of Polit. Econ.*, Oct. 1956, *64*, 416–24.

Political Constraints on Contractual Redistribution

By James M. Buchanan and Winston C. Bush*

Modern contract "theories" of distribution represent significant advances over the value-laden statements of preference that were passed off as serious intellectual constructions by the utilitarians. The modern theories attempt to "explain" observed institutions on the basis of conceptual contractual agreement among members of defined political communities. These theories may be classified in three distinct sets: preconstitutional, constitutional, and postconstitutional.

In the first, a preconstitutional state of anarchy is postulated, and hypotheses are then derived concerning the types of property rights that might emerge. Given a property rights structure and assumptions concerning individual behavior, conjectures can be made concerning the distribution of property. (See Buchanan 1973 or Bush.) In the second set, the contractual process occurs at the constitutional level where the position of individuals (families) is not fully identifiable; here the alternatives for choice are institutional arrangements that are presumed to remain in being over a succession of time periods. In the third set, the contractual process occurs postconstitutionally, within a defined institutional framework, and the alternatives for choice are explicit transfers of income and wealth among individuals and groups in the comunity. In the second or constitutional category, we place John Rawls's difference principle and also the related insurance principle.[1] In the third set, we place the so-called Pareto-optimal redistribution models, based on utility interdependence.[2] As a somewhat in-between model, we note the self-protection theory recently advanced by G. Brennan.

We restrict our discussion to contractually derived institutions that might emerge at the constitutional stage. We postulate that individuals are wholly uncertain about their prospective income-wealth positions in the periods for which the institutional structure to be chosen is to be applicable. Individuals make their own decisions, each behind a Rawlsian "veil of ignorance." For our purposes, this is a positive assumption about the actual states of persons making constitutional choice; it is not a normative statement concerning how persons should conceive themselves in making choices.

We want to address ourselves specifically to the questions raised by the recognition, at the time of constitutional contract, that actual in-period transfers of income and/or wealth must be implemented within an institutional setting peopled by individuals whose income-wealth positions are known. In other words, our questions concern the potential viability of the "terms of constitutional contract," the difficulties that might arise in enforcing these terms in a practically working political process, and, importantly, the possible feedback effects that these latter consider-

---

* Center for Study of Public Choice and Department of Economics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia. The final draft of this paper was submitted before the tragic death of Professor Bush on December 1, 1973.

[1] See Rawls; also, Buchanan and Gordon Tullock.

[2] See Harold M. Hochman and James D. Rogers.

TABLE 1

Individual	1	2
No-Transfer Income Distribution		
A_1	$10,000	$10,000
A_2	10,000	10,000
A_3	1,000	10,000
A_4	1,000	1,000
A_5	1,000	1,000
Posttransfer Income Distributions		
A_1	$ 5,000	$ 6,333
A_2	5,000	6,333
A_3	4,000	6,333
A_4	4,000	6,000
A_5	4,000	6,000

ations may exert on the constitutional decision itself. These are issues which have not, to our knowledge, been examined in the burgeoning modern literature on contractual distribution.

We define a potentially viable constitutional contract as one in which a majority of the individuals in the postconstitutional stage benefit when the terms of the contract are executed. Although the relationship between viability and majority rule is arbitrary, it does not substantially affect the results, and also conforms with the orthodox idealizations of the institutions of democracies.

Let us assume, first, that individuals in a constitutional stage reach agreement on an institutional structure that will generate in-period transfers from the relatively rich to the relatively poor, as measured against the no-transfer distribution that would emerge from pure market payments to resource owners under the constitutionally defined property rights in existence.[3] Initially, we assume that there is only one period in the postconstitutional sequence. Assume more specifically that the contractual agreement dictates that posttransfer incomes of the lowest income recipients shall be maximized in accordance with some Rawlsian-like difference principle. Table 1 presents two separate no-transfer and posttransfer distributions, for a community of five persons. (Note that measurable and adverse incentive effects are incorporated in these examples.)

As Table 1 suggests, the enforceability of redistribution schemes will depend significantly on the pattern of pretransfer market distribution. In Distribution 1, the assumed contractual rule for income transfers would clearly be viable. A majority of the community's members would benefit from it; indeed, by definition, no other rule could be better for them. By sharp contrast, in Distribution 2 implementation of the constitutional rule for redistribution will require that a majority of the community's members suffer transfer losses. In the absence of direct utility interdependence, implementation of the previously agreed-on constitutional rule for redistribution will be difficult to secure.

We may now consider the obverse of the above example. Suppose that there is no agreement on transfer policy reached in the constitutional state, which is equivalent to saying that the market-determined shares are to be left unchanged. In this instance, note that the results may be viable only if Distribution 2 emerges. If Distribution 1 emerges, postconstitutional transfers will tend to emerge, regardless of the absence of constitutional agreement.

The single-period arithmetical example is extremely simple, but it demonstrates one central point. The enforceability and maintenance of any constitutionally determined institution for income-wealth distribution will depend on the actual pattern of pretransfer distribution. The individual's expectations about this pattern will, therefore, influence his choice among alternative constitutional rules on income-wealth transfers. Other things equal, he will be more egalitarian in his constitutional choice the larger the proportion of relatively poor members he expects to be in the pretransfer pattern of distribution since this implies an increased probability that he himself will be poor. Note particularly that his result depends only on the relative numbers of poor and rich and not at all on the relative income levels of the two groups.

---

[3] For purposes of this discussion, we are restricting the analysis to the "distribution branch" of government budgets, in Richard Musgrave's terminology.

We can now relax the single-period assumption and examine the applicability of the chosen constitutional rule over a sequence of periods. We assume, as before, that at the time of constitutional choice the individual is wholly uncertain as to his own position in the market distribution that will emerge in the first period of the postconstitutional sequence. We need to specify, however, the relationship of the distributional position of an individual in each of the several postconstitutional periods. If his position is the same in all periods, the multiperiod model is no different from the single-period one. At the opposite extreme, assume that individual positions in the separate distributions are wholly unrelated. The individual who finds himself in a favorable income position in the first period remains wholly uncertain about his position in the second, etc. Consider the model in which Distribution 1 is predicted to emerge in each of the several periods of the sequence. As suggested, the Rawlsian redistributional rule is viable in the one-period setting. It will continue to be viable in the multiperiod setting under these assumptions only if individuals maintain the same set of attitudes or tastes for redistribution policy that they held at the time of the initial constitutional contract. If individuals observe that their own positions are unrelated over separate periods, and that these shift among different income levels, this experience may itself modify attitudes. Individual members of a political majority, in a single period, may find themselves unwilling fully to implement the redistribution that is dictated by the Rawlsian constitutional precept. In more general and relevant terms, this is the effect of observed upward social mobility on attitudes toward redistributional alternatives. Individuals may be unwilling to transfer maximinal incomes (dictated by Rawlsian precepts) from the relatively rich to themselves if they expect that any policy, once introduced, will become permanent.

The introduction of a multiperiod sequence reduces somewhat the expected political viability of the extreme distributional rule postulated under the predictions that Distribution 1 will prevail in all periods. But the basic conclusion that a less extreme version of this scheme might well be selected and expected to be viable holds.

Consider now the prediction that Distribution 2 will characterize market results over the whole sequence of postconstitutional periods. As indicated, the redistributional alternative would not be expected to be maintained in the one-period model, and a no-transfer policy might be the only agreed-on rule. In the multiperiod model, however, something between these extremes might emerge. Individuals who are in the relatively rich majority coalition, but who are wholly uncertain as to their own positions in subsequent income periods, will have some incentive to build in transfer mechanisms on the expectation that, once established, these policies will tend to be accepted in subsequent periods.[4] To accomplish this, they may accept net transfer losses during the period when income shares are known. In anticipation of this, there may be constitutional agreement on a redistributive scheme, even in the expectation that Distribution 2 will characterize each income period, an agreement that will be expected to be implemented by the majority coalition in the community.

---

[4] The argument here is related to Peter Hammond's.

The general point made with respect to the single-period model remains valid, however. More redistribution will be constitutionally agreed upon if Distribution 1 is predicted than if Distribution 2 is predicted. The principle may be more generally stated. If *all* members of a community agree on a redistributional scheme at the constitutional level, those persons who find their income positions settled even for a single period will find their postconstitutional preferences for redistribution inversely related to income position. In the multiperiod model, lifting of the "veil of ignorance" for even a single period will modify the costs and benefits of any previously agreed redistribution scheme.

In real-world circumstances, something between the two extreme models of income expectations surely prevails. An individual's income-wealth position in any one period is related to that which he expects to hold in subsequent periods, although there may well exist considerable uncertainty. In this setting, results that fall somewhere between the single period and the multiperiod models can be derived.

The examples have remained unreal because of the assumption of two distinct income classes. Somewhat more realism may be introduced by adding a middle class of income recipients in an example that becomes slightly more descriptive of existing distributional patterns. Consider a nine-member community, with no-transfer or market incomes arrayed as depicted in Table 2, Column 2. Assume that this pattern of market distribution is predicted to emerge in each postconstitutional period, but that the individual remains wholly uncertain as to which position he will come to occupy as the actual pattern emerges.

Initially, assume that there will be only one postconstitutional period, or, what amounts to the same thing, that an individual's position in the distribution will remain unchanged once it is settled in period one. From Table 2 it is clear that if there are no incentive effects, a policy of full equalization will be politically viable. Since the mean income is above the median, at least a majority of the persons in the group will find themselves improved as a result of the constitutionally selected redistribution policy. If incentive effects are allowed to enter the model, however, this conclusion need not follow. The set of transfers that would be required to maximize the position of the least-advantaged person in the predicted distributional array need not be consistent with the political structure.

Suppose for example that, ignoring incentive effects, a policy of full posttransfer equalization is adopted in the constitutional assembly. As individuals earn incomes in the market, however, they will be motivated in part by the knowledge of this equalization policy. As a result total community income falls to such a level that the income realized by each person amounts to, say, only $3,500, as opposed to an average income of $5,167 when no transfers are anticipated.

The position of the least advantaged may be improved if some allowance of inequality is reintroduced. If a benevolent despot is controlling, the position of the

TABLE 2

Individual	Predicted No-Transfer Distribution	Idealized Rawlsian Rule
A_1	$10,000	7,000
A_2	8,000	6,500
A_3	6,000	5,500
A_4	5,500	5,200
A_5	5,000	4,600
A_6	4,500	4,400
A_7	4,000	4,000
A_8	1,500	4,000
A_9	1,000	4,000
Mean Income	$ 5,167	
Median	5,000	

least advantaged may, say, be increased to an income level of $4,000, with a posttransfer income array like that indicated in Column 3, Table 2. Note, however, that this posttransfer distribution would not be supported in a strictly democratic political process. Six of the nine members of the community find their own posttransfer incomes to be lower than those which they secure in the no-transfer setting.

The example suggests that the minimal political constraint on the implementation of any redistributional rule would be the requirement that at least one-half of the community's members find their own income positions improved over the no-transfer setting. Further consideration of the limits that political process may place on redistribution rules may suggest more restrictive constraints. Evidence suggests that the observed beneficiaries of income transfers in democratic structures are middle-income rather than low-income recipients, and public-choice arguments have been advanced to explain this observation.[5] As a plausible constraint reflecting something of this, models might be examined which require that the median members of the community receive absolute transfers at least as large as persons falling lower on the income scale.[6]

Our purpose here is not to examine alternative constraints in detail. We make the more general point that political constraints must be recognized to exist when the problem of implementing constitutionally approved rules on income-wealth redistribution is discussed. In any real-world setting, of course, the discussion of institutional rules affecting income-wealth distribution must take place in recognition of existing legal definitions of property rights, of existing political decision-making mechanisms, and of predicted patterns of income distribution as well as predicted positions of persons within these predicted distributions.

---

[5] See George Stigler and Tullock.

[6] See Buchanan (1972).

REFERENCES

G. Brennan, "Pareto Desirable Redistribution: The Non-Altruistic Dimension," *Public Choice*, Spring 1973, *14*, 43–68.

J. M. Buchanan, "The Political Economy of the Welfare State," paper presented at Capitalism and Freedom Conference, Charlottesville, October 1972.

———, *The Limits of Liberty: Between Anarchy and Leviathan*, unpublished monograph, Blacksburg 1973.

——— and G. Tullock, *The Calculus of Consent*, Ann Arbor 1962.

W. C. Bush, "Some Implication of Anarchy for the Distribution of Property," *J. Econ. Theory*, forthcoming.

P. Hammond, "Charity: Altruism or Cooperative Egoism," paper presented at Russell Sage Foundation Conference, New York, March 1972.

H. Hochman and J. Rodgers, "Pareto Optimal Redistribution," *Amer. Econ. Rev.*, Sept. 1969, *59*, 542–547.

R. A. Musgrave, *Theory of Public Finance*, New York 1959.

J. Rawls, *A Theory of Justice*, Cambridge, Mass., 1971.

G. J. Stigler, "Director's Law of Public Income Redistribution," *Jour. of Law and Econ.*, April 1970, *13*, 1–10.

G. Tullock, *The Social Dilemma*, unpublished monograph, Blacksburg 1973.

DISCUSSION

Alvin K. Klevorick, Yale University: I shall use John Rawls's remarks on the maximin criterion as the organizational framework for my comments. In particular, I shall focus on his more difficult comparison—the one between his two principles of justice and a variant of these in which the difference principle is replaced by the principle that "social and economic inequalities are to be adjusted so as to maximize average utility consistent with fair equality of opportunity." Rawls presents several reasons which, he believes, support the maximin criterion in this comparison. I would like to question whether, upon further examination, these arguments provide as strong a case as he suggests they do. The paper by Dennis Mueller and the one by James Buchanan and Winston Bush can be viewed as bearing in important, though different, ways on the arguments Rawls puts forth in making his comparison.

Rawls's first argument is based upon the "very considerable normal risk aversion" individuals will display in the original position. He grants that, because of the crucial nature of the choice faced in that position, the utility criterion may not differ much from maximin, and he cites the result that the standard of utility approaches maximin as risk aversion increases without limit. Nevertheless, he writes, "I assume that, based on considerations of risk aversion alone, there is a significant difference between the two criteria." But that is all the risk aversion argument comes to: an assumption.

Second, Rawls argues that "much less information is needed to apply the maximin criterion." While maximin may require less information than average utility maximization, I think the difference is smaller than Rawls suggests. First, the same group may not be least favored under each possible arrangement so that under maximin one has the considerable task of determining the least favored group under all alternative arrangements. Second, this task requires one to measure the utility of each group and compare the utility of different groups to determine which one is the least favored under a particular arrangement. Third, the method of estimating the overall balance of gains and losses in the average utility setting is readily at hand since it has been agreed that the goal is maximization of the *average* of the utilities. Rawls states that in applications the average utility principle "leaves so much to judgment that some may reasonably claim that the gains of one group outweigh the losses of another, while others may equally reasonably deny it." It would seem that such reasonable differences of opinion could also arise under maximin in determining which is the least favored group. This suggests that a real cost under both of the two sets of principles Rawls is comparing is the cost of resolving such differences of opinion. Or does his statement that everyone reasons in the same way exclude this problem? If it does, then surely it must also apply to resolving the disputes to which he alludes in the average utilitarian context.

This brings me to the largest part of the information required in the contractarian setting being considered—information needed for both maximin and the average utilitarian rule. It is the information to which Mueller devotes his discussion—knowledge which enables the individual "to place himself in the position of those actually or potentially subject to the rules . . . knowledge of those with whom he might exchange places." Mueller discusses the costs of achieving such knowledge, the tradeoff between acquiring knowledge at the intensive and knowledge at the extensive margin, and the fact that formation of the polity does not necessarily directly contribute to the kind of knowledge and understanding needed for a just society. The fact that this major information requirement is common to Rawls's two principles and the variant with which he is comparing them also narrows the difference between the informational demands of the two approaches.

Having mentioned Mueller's ideas, let me digress for a moment to consider the paradox with which he left us at the end of his paper. I think it is a paradox only because he overstates the necessary conditions for a just society. The developments he points to —technological change, specialization, mobility, increasing education, urbanization— will tend to produce the knowledge required for the just society because they tend to reduce the differences among individuals. But they lead to only one kind of just society. Surely, there is a variety of just societies, and the just society Mueller—and most of us—would find more interesting is one where differences exist and are appreciated and understood. To be sure, in terms of the cost of acquiring the information Mueller discusses, achieving the "interesting just society" will be more expensive than achieving the "dull just society," but not all just societies need be marked by "loneliness and indifference" as Mueller seems to fear.

Let us now return from our brief digression to Rawls's case for maximin. His third argument is that maximin is more suitable as a public principle. The maximin principle is sharper than the average utilitarian principle: one would expect maximin to be understood better by citizens generally. But, while it may be difficult for the public to verify that average utility has been maximized, one should not understate the difficulty the public would have in verifying that the least favored group's position has been maximized. In particular, preliminary to any such verification would be the need for all members of the public to agree about exactly which is the least favored group. I have already questioned whether such agreement will be readily achieved.

The next reason Rawls offers in support of the maximin criterion is that it imposes weaker strains of commitment: "in the original position the parties are to favor those principles compliance with which should prove more tolerable, whatever their situation in society turns out to be." He discusses the difficult strains which face the less advantaged under the utility principle. In contrast, notes Rawls, under maximin the serious strains of commitment are faced by the better situated. Just as the less advantaged must accept less under the utility principle than they could have otherwise (for example, under maximin), so the more advantaged must settle for less under maximin than they could receive otherwise (for example, under the utility principle). Two factors are, however, supposed to lessen greatly the strains those who are better off face under maximin: (a) they are more fortunate and enjoy the benefits of that fact and (b) insofar as they value their position relative to others', they give up that much less. There would seem to be a problem with each of these supposedly mitigating factors. Argument (a) is implicitly based on the assumption of diminishing marginal utility of income. Otherwise, why should we count the dissatisfaction of the better off less than the dissatisfaction of the less well-situated? Argument (b) is based on an assumption about human behavior which, though it may be confirmed by armchair (and some careful) empiricism, introduces the very kind of strong interdependence of utility functions that Rawls excludes his individuals from having or considering in the original position.

The Buchanan and Bush paper can be viewed as a positive statement about the strains of commitment under alternative constitutional contracts. Their general point is that "political constraints must be recognized to exist when the problem of implementing constitutionally-approved rules on income-wealth redistribution is discussed." Essentially, Buchanan and Bush consider the possibility that individuals in a post-constitutional stage might break contracts they had made at the constitutional stage. That is, a potential constitutional contract might prove to be nonviable because a majority of individuals would lose in the post-constitutional stage if the terms of the contract were executed. Rawls excludes the possibility that constitutional contracts might be broken. Nevertheless, the Buchanan and Bush examples, which show the difficulties that arise in enforcing and maintaining constitutionally-determined institutions for income-wealth redistribution in a practically-working political process in the postconstitutional world, can be interpreted as illustrating the strains of commitment Rawls discusses. And the Buchanan and Bush examples show, specifically, how difficult those strains may be when Rawls's maximin principle is used as the basis for redistribution.

There are two specific points I would like to raise about the Buchanan and Bush paper—the first is a question, the second is a comment. First, in all their examples, the individual makes a single-point prediction about the pretransfer income distribution; he thinks in terms of one distribution emerging. What happens if the individual thinks either Distribution 1 or Distribution 2 could emerge, with probabilities p and $1-p$, respectively? (One could obviously go on to a subjective distribution over a larger number of income distributions.) How does this alter the individual's behavior at the constitutional level? Buchanan and Bush say that "The individual will, other things equal, be more egalitarian in his constitutional choice the larger the proportion of relatively poor members he expects to be in the pretransfer pattern of distribution." How does this result get modified in the case of many possible pretransfer income distributions? Does the individual look at the weighted proportion of relatively poor people?

The second point concerns a conclusion Buchanan and Bush draw in their discussion of the multiperiod case: "If *all* members of a community agree on a redistributional scheme at the constitutional level, those persons who find their income positions settled even for a single period will find their post-constitutional preferences for redistribution inversely related to income position." Since the "draws" for positions in the pretransfer income distribution in successive periods are totally independent in the examples Buchanan and Bush use to generate their multiperiod results, the probability that an individual will find himself in a particular income class is the same in each period, and it is independent of where he has been in the past. Hence, unless individuals in the Buchanan and Bush world do not understand the probability calculus—for example, they systematically underestimate the probability of runs—the inverse relation between redistribution preferences and revealed income class must be due either to changes in individual utility functions through time or to differences in the degree of risk aversion at different income (wealth) levels. No explanation is provided for the former type of systematic change in the utility functions through time. On the other hand, if changes in the degree of risk aversion with wealth are the source of the inverse relation in question, then it is interesting that what is required is increasing absolute risk aversion. This is precisely the opposite of what one usually assumes about individuals when discussing their behavior as investors and portfolio selectors; they are usually assumed to show decreasing absolute risk aversion.

The last argument Rawls puts forth in support of the maximin criterion is the one he finds most compelling. He writes, "I want to suggest that the aspirations of free and equal personality point directly to the maximin criterion." As Rawls indicates, when his two principles are followed, the natural distribution of abilities is viewed in some ways as a collective asset. Furthermore, the maximin criterion would conform to the rule "from each according to his abilities, to each according to his needs" if society were to impose a lump-sum-tax-and-transfer program based on ability levels. Rawls finds two major problems with such a lump-sum tax scheme. First, there are considerable practical difficulties involved in operating such a program. Second, and more important, the scheme would involve "a drastic infringement with freedom"—"greater talents are not a collective asset in the sense that society should *compel* those who have them to put them to work for the less favored" (emphasis added). "But," says Rawls, "society can say that the better endowed may improve their situation only on terms that help others."

Rawls has set out here—in very cogent fashion—a fundamental tension in the post-constitutional life of the society he discusses: the tension between freedom of behavior and *ex post* equality. A preliminary question that must be confronted is how should one measure the degree of compulsion or degree of interference with liberty. Clearly, any tax system interferes with liberty to some degree; any such system involves some compulsion. But presumably tax programs which allow people to choose, for example, between market (taxable) and nonmarket (nontaxable) uses of their natural assets (for example, time and abilities) would be considered less interfering or less compulsive than tax systems which are structured so as to disallow such choices. Bypassing that preliminary question in this discussion, one can still ask why maximin provides the best resolution of the liberty-equality tension and what "best" means in this context. Why, for example, is maximin (which chooses a particular liberty-equality mix) superior to a rule which says that the better endowed may improve their own situation only if they improve the situation of the least advantaged by x percent—or by at least as much as they improve their own situation? Put another way, why does maximin provide the optimal degree of coercion in a Rawlsian world?

NEW DIMENSIONS OF ECONOMIC INEQUALITY

The Concentration of Personal Wealth, 1922-1969

By JAMES D. SMITH AND STEPHEN D. FRANKLIN*

This paper presents estimates of the concentration of personal wealth in the United States from 1922 to 1969. These estimates lead us to conclude that the distribution of wealth (1) became significantly more equal in the 1930's and early 1940's, two periods of massive government intervention in the marketplace, and (2) has remained essentially unchanged since 1945.[1]

In what follows, we compare the wealth held by the richest 1.0 and 0.5 percent of the population to that of all persons. The wealth of the richest 1.0 and 0.5 percent was estimated by the estate multiplier technique.[2] The wealth of all persons was derived from national balance sheets.[3]

The estimates presented here for the period before 1953 were developed by Robert J. Lampman using highly aggregated Internal Revenue Service (*IRS*) data. For 1953 and 1958 we use detailed estimates (from special *IRS* tabulations) by Lampman and Smith, modified slightly to take account of current knowledge. Estimates for 1962, 1965, and 1969 are new detailed estimates developed by the authors using microdata files of estate tax returns prepared by the *IRS* for its routine publications. Here we focus on the years since 1953.

Information available from estate tax returns varies from year to year, so a number of adjustments were made to bring the estimates for individual years into conceptual alignment with one another. The alignment problem was exacerbated because the *IRS* has destroyed tapes of returns filed before 1963, leaving only Lampman's and Smith's printed tabulations for 1953 and 1958.[4] It was impractical to reestimate the distributions for 1953 and 1958 by better methods based on current knowledge. Consequently, the estimates for 1962, 1965, and 1969 were made consistent with those for 1953 and 1958

---

* The Urban Institute and the Pennsylvania State University. The work reported here is part of the Urban Institute's research program on income and wealth distribution. The support of the National Science Foundation is gratefully acknowledged.

[1] We wish to make clear that our concern is with temporal change and that we have sacrificed best estimates for individual years to achieve consistency over the time series. Individual figures have a downward bias of 10 to 15 percent from our best estimates of concentration. (Best estimates for 1969 may be found in Smith; similar estimates for other years will appear later.)

[2] Detailed descriptions of the methodology and attendant problems can be found in Smith and Staunton Calvert, Robert J. Lampman, and Smith.

[3] National balance sheets were constructed for a person's sector using data supplied by the Board of Governors of the Federal Reserve System. Helen Stone Tice did the basic work on these special sector balance sheets. Smith provides a detailed description of the balance sheet.

[4] A further problem resulted from the fact that the *IRS* erased the age field from the 1965 tape. This was most unfortunate because of all years for which the *IRS* has coded estate tax returns, 1965 had the most detailed classification of information. The erased data was restored by a stochastic process which took into account the relationship between age and other characteristics observable in the files for 1962 and 1969.

after they were slightly adjusted.[5] The result is a time series of wealth distributions which understates the degree of concentration in any given year but is consistent and permits comparisons over time.

As applied here, the estate multiplier technique treats decedents in a given year as a stratified sample of the living population whose estates would have been required to file tax returns had they died during the year. As with any sample, estimates of population parameters can be made by weighting observations by the reciprocals of their sampling rates. For samples "drawn" by death, the sampling rates are mortality rates. The mortality strata used here are age, sex, and social class.

We first look at the overall concentration of wealth from 1922 to 1969 and at the detailed estimates for 1953 to 1969, and then discuss briefly the estimation of trust assets.

I. Changes in the General Concentration of Wealth, 1922–1969

Wealth in the United States has become less concentrated in the last half century. The diminution has not been great, however, and it all occurred during periods when the market system was functioning under duress or was in administrative abeyance, specifically, the Great Depression and World War II.

Lampman, in his classic study of the distribution of wealth, found the highest measurable concentration on the eve of the Great Depression. According to his figures, the richest 0.5 percent of Americans in 1929 owned 32.4 percent of the net wealth of all individuals, no doubt a reflection of the phenomenal rise in the price of stock—long the dominant asset in the portfolios of the rich—at that time.

But the collapse of the market in October of 1929 and its plummet downward to 1932 carried with it the fortunes of many near the pinnacle of America's wealth pyramid. Whether 1932 was the nadir of America's affluent we do not know, but by 1933 the top 0.5 percent had lost 22 percent of the share of wealth it held four years before. Lampman found that the share of the richest 0.5 percent then rose somewhat—from 25.2 percent in 1933 to 28.0 percent in 1939—but later fell sharply to only 20.9 percent in 1945. And 1949 saw the most egalitarian distribution of American wealth measured to date. In that year the top 0.5 percent of wealth-holders held only 19.3 percent of all personally held net wealth.

By 1953, Lampman found, the share of the top 0.5 percent had risen sharply to 22.7 percent (22.0 percent after our subsequent adjustments). His rough estimate for 1954, based on published data, showed about the same share, 22.5 percent. His estimate for 1956, also based on published data, showed a substantial increase to 25.0 percent.

Following Lampman, Smith estimated that the top 0.5 percent held 21.7 percent of the nation's personal wealth in 1958, a figure not significantly different from Lampman's for 1953. And *IRS* estimates imply that the top 0.5 percent held 23.4 percent of the wealth of all persons in 1962.[6]

---

[5] We adjusted Lampman's basic estimates for 1953 to include an amount represented by decedents whose age was unspecified on estate tax returns. Smith's estimate for 1958 was distorted by a sampling fluke—the inclusion of an outlier, i.e., a totally atypical individual (with a very large estate held entirely in annuities). We therefore adjusted his figure by distributing this outlier's assets. (Had the outlier's assets been left unadjusted, the estimate of the total value of annuities would have been raised by over $16 billion.)

[6] We believe, on the basis of information provided us by the Metropolitan Life Insurance Company, that the mortality rates used for the 1962 estimates resulted in a significant overstatement of the wealth of persons with over $60,000 gross assets.

We also believe that preliminary *IRS* estimates for

It might seem, then, that the trend toward equality in the two decades following the market's crash in 1929 reversed itself after World War II. But our examination of the old evidence and our new estimates for 1958, 1962, 1965, and 1969 do not support this conclusion. In our view, observable point-to-point differences in the estimated concentration of wealth since 1945 are due to inconsistencies and inadequacies in *IRS* data, sampling errors, short-run stock market variations, and a valuation bias in the basic data which understates the wealth of the rich in declining market periods by permitting estates to select between the most beneficial of two valuation dates.

In Figure 1 we have plotted estimates of the share of the nation's personal net worth held by the richest 0.5 percent of all individuals since 1922, along with Standard and Poor's stock price index. One can perceive a generally downward trend in concentration up to the mid-1940's. After that it appears that the trend line is flat. It also appears that, after 1945, periods when the actual wealth share was above (below) the trend were generally preceded by periods of market increases (declines).

II. The Detailed Concentration of Wealth, 1953–1969

Table 1 gives estimates of the shares of specific assets held by the richest 0.5 percent and richest 1.0 percent of the population in 1953, 1958, 1962, 1965, and 1969.

The only striking change over this period was a decline in the concentration of corporate stockholdings. The richest 0.5 percent of the population held 86 percent of the value of personally owned stock in 1953, but only 44 percent in 1969.[7]

It is worth noting that 1965 is the first year for which an estimate of the concentration of trust assets by size of individuals' wealth is available. Although the estimates for other years are derived estimates based on the 1965 measurement and a procedure described below, the 1965 estimate is a direct one. It is clear that trusts are the most concentrated of all substantial types of wealth: 85 percent of the value of trusts were in the hands of the top 0.5 percent of wealth-holders and 93 percent were held by the richest 1.0 percent in 1965.

Before discussing the estimation of trust assets, we once again wish to emphasize that the values shown here understate shares of wealth owned by the top 0.5 percent of wealth-holders because maintaining consistency with the published data for 1953 and 1958 prevented the application of refinements usable on the 1962, 1965, and 1969 microdata.

III. Estimation of Trust Assets

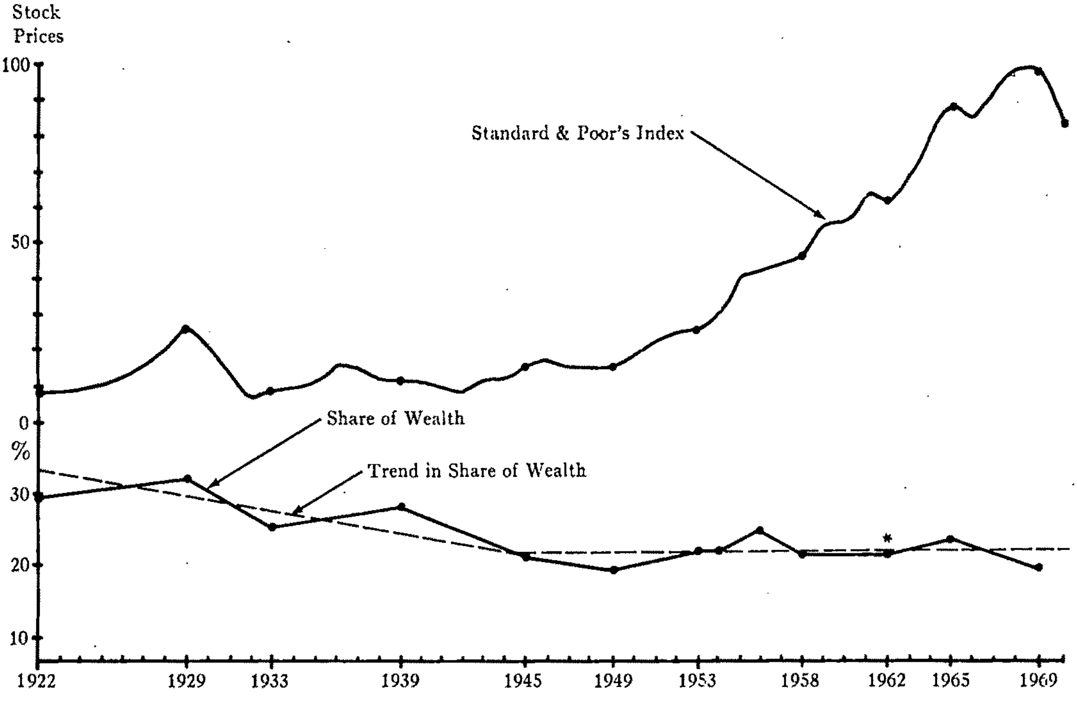
In 1962 and 1969 the *IRS* recorded trust assets by type of property (real estate, stock, etc.), but not a separate total for trusts. In 1965 it did exactly the reverse. The data for 1953 and 1958 included trust interests in "other property." Since the value of trust assets held by the wealthy is substantial, we attempted to make the estimates consistent across years by estimating the total value of trust interests of the top 0.5 percent and 1.0 percent of the population in those years where trusts are included in the "other property" category and then allocating this total among specific asset types (stocks, bonds, etc.), based on the composition of trust assets in national balance sheets.

The value of trust assets reportable on estate tax returns is an actuarial value. It

---

1969 (see Vito Natrella) are too high. They imply that persons with gross assets over $60,000 owned 47.4 percent of total wealth in 1969; adjusted to reflect underestimation biases, the implied share of such persons will exceed 50 percent.

[7] This is consistent with New York Stock Exchange figures showing an increase of nearly 500 percent in the number of shareholders between 1952 and 1970.



* *IRS estimate.*

FIGURE 1—SHARE OF WEALTH HELD BY RICHEST 0.5 PERCENT OF POPULATION, AND STOCK PRICES, 1922 TO 1969

is, in general, the present value of future payments which may be contingent upon such events as the death of another person. Although the valuation of trust assets in this manner may be appropriate for estate tax purposes, it understates the immediate market value of trust assets in which persons have a financial interest. Consequently, the value of trust assets estimated by the estate multiplier technique cannot be compared directly to national balance sheet totals, which represent the market value of trust assets, without regard to restrictions imposed by trust instruments. To determine the extent to which the value of trust assets is understated, we estimated the number of trusts valued at over $60,000 by the estate multiplier method using the 1965 file (the one file on which they are identified). The estimated number of trusts over $60,000 was 118,925. Next, the number of such trusts was counted off, starting with the largest, in the array of fiduciary income tax returns filed for trusts in 1965.[8] The reported income of the counted-off trusts was then capitalized at the rate which the reported income of all trusts bore to the value of trust assets in the 1965 national balance sheets, or about 5.7 percent. The capitalized income of these 118,925 largest trusts amounted to $94.9 billion. The trust assets estimated by the estate multiplier method for the same number of largest trusts in 1965 was $51.5 billion. On this basis, we concluded that conceptually we were limited to estimating only 54.3 percent of balance sheet trust assets. Consequently, we reduced the value in the national balance sheet to 54.3 percent *for purposes of*

---

[8] See U.S. Department of Treasury, p. 22.

TABLE 1—SHARES OF RICHEST[a] 0.5 PERCENT AND 1.0 PERCENT OF PERSONS IN NATIONAL WEALTH, 1953, 1958, 1962, 1965 AND 1969

Asset	1953					1958				
	Value Held by Richest			Share Held by Richest		Value Held by Richest			Share Held by Richest	
	100.0%	0.5%	1.0%	0.5%	1.0%	100.0%	0.5%	1.0%	0.5%	1.0%
		billions		*percent*			*billions*		*percent*	
Real estate[b]	$ 439.0	$ 45.0	$ 68.0	10.3	15.5	$ 621.5	$ 62.5	$ 93.9	10.1	15.1
Corporate stock[c]	151.5	116.6	130.8	77.0	86.3	264.1	175.9	199.2	66.6	75.4
Bonds	72.8	33.0	8.3	45.3	52.6	87.0	31.3	36.0	36.0	41.4
Cash[d]	160.1	20.9	28.8	13.1	18.0	216.0	22.5	32.8	10.4	15.2
Debt instruments[e]	34.0	8.2	10.9	24.1	32.1	43.7	12.5	16.3	28.6	37.3
Life insurance (*CSV*)[f]	64.5	6.6	9.1	10.2	14.1	79.9	7.5	11.3	9.4	14.1
Miscellaneous and trusts[g]	243.3	30.0	38.6	12.3	15.9	343.2	45.6	52.8	13.3	15.4
Trusts	20.5	17.5	18.8	85.4	91.7	30.3	25.8	27.9	85.1	92.1
Miscellaneous	222.8	12.5	19.8	5.6	8.9	312.9	19.8	24.9	6.3	7.9
Total assets	$1144.7	$242.8	$305.7	21.2	26.7	$1625.1	$332.0	$414.4	20.4	25.5
Liabilities	$ 140.0	$ 21.3	$ 29.0	15.2	20.7	$ 227.4	$ 29.2	$ 38.3	12.9	16.8
Net worth	$1004.7	$221.5	$276.7	22.0	27.5	$1396.7	$302.8	$376.1	21.7	26.9
		millions					*millions*			
Number of persons		0.80	1.60				0.87	1.74		

Asset	1962					1965				
	Value Held by Richest			Share Held by Richest		Value Held by Richest			Share Held by Richest	
	100.0%	0.5%	1.0%	0.5%	1.0%	100.0%	0.5%	1.0%	0.5%	1.0%
		billions		*percent*			*billions*		*percent*	
Real estate[b]	$ 770.0	$ 79.6	$117.8	10.3	15.3	$ 917.7	$ 94.4	$135.8	10.3	14.8
Corporate stock[c]	426.4	227.3	264.4	53.3	62.0	596.6	317.2	364.9	53.2	61.2
Bonds	94.5	33.2	38.4	35.1	40.6	103.6	57.5	63.2	55.5	61.0
Cash[d]	278.3	28.9	42.5	10.4	15.3	366.0	43.7	62.7	11.9	17.1
Debt instruments[e]	51.5	16.5	21.8	32.0	42.3	53.3	19.8	25.4	37.1	47.7
Life insurance (*CSV*)[f]	93.8	7.1	10.7	7.6	11.4	107.2	6.5	10.9	6.1	10.2
Miscellaneous and trusts[g]	425.5	NA	NA	—	—	514.1	85.3	101.8	16.6	19.8
Trusts	46.1	NA	NA	—	—	57.5	49.0	52.7	85.2	91.7
Miscellaneous	379.4	39.8	52.7	10.5	13.9	456.6	36.3	49.1	8.0	10.8
Total assets	$2093.9	$432.4	$548.3	20.7	26.2	$2601.0	$575.4	$712.7	22.1	27.4
Liabilities	$ 314.0	$ 47.8	$ 61.0	15.2	19.4	$ 413.3	$ 57.0	$ 73.1	13.8	17.7
Net worth	$1779.9	$384.6	$487.3	21.6	27.4	$2187.7	$518.4	$639.6	23.7	29.2
		millions					*millions*			
Number of persons		.93	1.87				.97	1.94		

Asset	1969				
	Value Held by Richest			Share Held by Richest	
	100.0%	0.5%	1.0%	0.5%	1.0%
		billions		*percent*	
Real estate[b]	$1188.8	$117.0	$170.7	9.8	14.4
Corporate stock[c]	832.5	366.3	423.3	44.0	50.8
Bonds	198.9	63.7	71.5	32.0	35.9
Cash[d]	495.0	48.1	71.2	9.7	14.4
Debt instruments[e]	85.3	21.9	29.6	25.7	34.7
Life insurance (*CSV*)[f]	127.2	8.4	13.8	6.6	10.8
Miscellaneous and trusts[g]	705.8	107.0	133.2	15.2	18.9
Trusts	70.4	60.0	64.5	85.2	91.6
Miscellaneous	633.7	47.0	68.7	7.4	10.8
Total assets	$3561.4	$672.4	$848.8	18.9	23.8
Liabilities	$ 557.5	$ 75.8	$100.5	13.6	18.0
Net worth	$3003.0	$596.7	$748.1	19.9	24.9
		millions			
Number of persons		1.01	2.03		

making estimates of the share of trust assets held by the richest 1.0 and 0.5 percent of the population.

Interests in trusts for the years 1953 and 1958 were estimated for the top 0.5 percent and 1.0 percent of the population using the information provided from 1965 returns. It was assumed that the share of trusts held by the wealthy was the same in each year as in 1965. After an estimate was made for trusts, the value was allocated to stocks, bonds, notes and mortgages, and cash in the same proportion which each asset represents of the total trust assets held by all persons. For the 1969 data, contrary to the *IRS* reports that trust assets were distributed by asset type, there is strong evidence that, in fact, the total value of trusts was included in "other property." The original estimates for 1969 showed that the share of stocks and bonds (the largest components of trust assets) had dropped to 39 percent and 25 percent respectively and that their share of miscellaneous property had risen to about 17 percent. The estimates for 1969 shown in Table 1 are based on the assumption that trust assets had not been allocated and were shown in "other property."

REFERENCES

R. J. Lampman, *The Share of Top Wealth-Holders in National Wealth,* Princeton 1962.

V. Natrella, "Discussion," in J. D. Smith, ed., *Personal Distributions of Income and Wealth,* forthcoming.

J. D. Smith, "The Concentration of Wealth in America," *Rev. of Income and Wealth,* forthcoming.

——— and S. Calvert, "Estimating the Wealth of Top Wealth-Holders from Estate-Tax Returns," *Proc. of the Bus. and Econ. Sec., Amer. Statist. Assn.,* 1965.

New York Stock Exchange, *Census of Shareholders,* New York 1972.

U.S. Dept. of Treasury, *Statistics of Income 1965, Fiduciary, Gift and Estate Tax Returns,* Washington 1966.

←

[a] "Richness" is measured in terms of gross assets. Net worth is preferred to gross assets as a classifier, but the microdata for 1953 and 1958 which would have permitted such an arrangement have been destroyed by the *IRS*. The microdata for 1962, 1965, and 1969 were therefore ordered by gross assets to produce estimates consistent with Lampman's and Smith's for 1953 and 1958 respectively.

[b] Real estate is shown at its market value without deduction of mortgages, liens, or other incumbrances. In 1953 and 1958 only real estate located in the United States is included. In 1962 the value of real estate located outside the United States was brought into the estimate by a change in the law which made foreign real estate subject to estate taxes. The amount of such real estate is, however, seriously underrepresented because the law took effect late in 1962. Only estates for decedents who died after October 16, 1962 and who had acquired foreign real estate (except by gift or inheritance) after February 1962 were required to report it on estate tax returns. In 1965 and 1969 foreign real estate was included along with other real estate.

Included in real estate are land and structures for personal and business use. All other business assets are included in the "miscellaneous" category. Real estate held in trust is included here to the extent of the trust interest. A relatively small proportion of trust assets is in real estate, but the absolute value of all trust assets is understated here for reasons explained in the text.

[c] Corporate stock includes the value of all common and preferred issues, shares in domestic or foreign firms whether traded or closely held, certificates and shares of building and loan and savings and loan associations, Federal Land Bank stock, and other instruments representing an equity interest in an enterprise. Accrued dividends are also included. Stock held in trust is included, but the absolute value is understated.

[d] Cash includes balances in checking and savings accounts, currency on hand or in safety deposit boxes, cash balances with stock brokers, and postal savings accounts. Cash in trust is included, but understated.

[e] Debt includes all legal obligations except loans on life insurance policies.

[f] Life insurance (cash surrender value—*CSV*) is the amount individuals could expect to receive were they to surrender their policies to the carriers. It takes account of policy loans, accrued dividends, and unearned premiums.

[g] "Miscellaneous and trusts" includes all assets owned in trust *except real estate* and all assets other than real estate, corporate stock, bonds, cash, debt instruments, and life insurance (*CSV*) not held in trusts. Included are such items as consumer durables, personal effects, business assets (excluding real estate), mineral rights, tax sale certificates, judgments, lifetime transfers, and growing crops if not included in the value of real estate. This classification is shown here as an information item to explicate certain adjustments described in the text. It should not be summed with other assets to arrive at a total asset figure because trust assets are included within the individual asset types.

Miscellaneous assets are those described under miscellaneous and trusts less trust assets. The miscellaneous asset category is added to other assets to arrive at total assets.

Trusts represent the actuarial value of reversionary and remainder interests in trusts. This actuarial value is substantially less than the total market value of assets held in trusts. On the basis of analysis reported in the text the national balance sheet totals (100 percent) have been adjusted to the reporting concept used for estate tax purposes.

The separate value of trusts could be estimated directly only for 1965. For other years indirect estimates were made by a method described in the text. The value for trusts is shown as an information item. The assets held in trust have been distributed to specific asset categories.

Who Paid the Taxes in 1966?

By Benjamin A. Okner and Joseph A. Pechman*

The effect of taxes on the distribution of income depends on the incidence of the major taxes in the tax system. We have just completed a new study on this subject, *Who Bears the Tax Burden?*, which will be published by the Brookings Institution in the spring of 1974. The study does not resolve the question of tax incidence but measures the difference in the distribution of tax burdens under several major competing views. We believe the study is of interest because of both its methodology and its substantive results.

I. The Data and Concepts

In the past, the income distribution data available for tax analysis have been deficient in two respects: first, they did not represent the entire income-receiving population; and second, they failed to include all the income known to have been received by that population. While annual information on income subject to tax is available from our federal individual tax returns, the omission of people not required to file distorts the distribution for those at the bottom of the income scale. The U.S. Census Bureau also collects income information each year from about 50,000 households which are a representative sample of the total civilian population. But, in addition to using a different population unit, the Census employs an income concept that excludes capital gains and other important income items. In addition, both the tax data and the Census data understate incomes in varying degrees and therefore cannot be linked with personal income and other aggregate statistical series.

To remedy these defects, we constructed a new file—called the MERGE file—that combines the best information available from both the Internal Revenue Service and the Census Bureau. In creating this file, we linked information on 30,000 families and single persons from the 1967 Survey of Economic Opportunity (*SEO*), conducted by the Census Bureau for the U.S. Office of Economic Opportunity, with data from a file containing information from 90,000 federal individual income tax returns filed for the year 1966.[1] All of this information is available on magnetic tape and can be processed quickly and efficiently on a high-speed electronic computer. The availability of the computer technology permitted us to prepare estimates in the type of detail that was never possible with the older data processing techniques.

Income. The income concept used in the study corresponds to an economist's comprehensive definition of income. In addition to the incomes earned in the production process—wages, interest, dividends, rents and royalties—it includes transfer payments and accrued capital gains. Gifts and bequests should also be included as income under this definition, but these were omitted because we know very little about their distribution among families.

This income concept corresponds to national income at factor prices (although

---

* Senior Fellow and Director of Economic Studies respectively, Economic Studies Division, The Brookings Institution. The views expressed are ours and not necessarily those of the officers, trustees, or other staff members of the Brookings Institution. The work described was financed by grants from the Ford Foundation and the U.S. Office of Economic Opportunity.

[1] For details on the methods used to create the file, see Okner.

national income, of course, excludes capital gains and transfer payments) but for our purposes it was necessary to adjust income shares to a market-prices basis. This required the inclusion of indirect taxes as well as direct taxes in income. Direct taxes, of course, are automatically included in factor shares. Indirect taxes were allocated to individual family units in proportion to their shares of factor incomes.[2] The resulting concept is called adjusted family income.

Taxes. The definition of taxes in the study is basically the same as that of government receipts, as defined in the national income accounts. However, since government receipts is a more comprehensive concept than taxes, nontax revenues were excluded from the tax measure.[3] In addition, tax collections attributed to institutions or organizations not in the household population were excluded, along with customs duties and estate and gift taxes. Thus, with minor exceptions, the procedure used was to accept the national income account definitions for the classification of tax and nontax receipts.[4]

Effective Tax Rates. Relative tax burdens in the study were measured by comparing effective rates of tax paid by family units. These were computed by expressing taxes as a percentage of adjusted family income, and thus reflect the proportion of the family's income which is accounted for by taxes. Although income is generally regarded as an acceptable measure of financial status, it should be noted that *income for a single year* (which may be unusually high or low for any of several transitory reasons) may be a poor indicator of "true" financial status for many families. Current-year income was used for measuring tax burdens in the study because income information for longer periods of time was not available.[5]

Incidence Assumptions. Since economists still disagree about the incidence of several of the most important taxes in the overall tax system and since we were able to use a computer to manipulate the MERGE file, we did not limit ourselves to any one view of tax incidence. Instead, we prepared estimates on the basis of eight different sets of incidence assumptions that span the range of opinions currently held by economists (see Table 1).

There is substantial agreement among economists on the incidence of the individual income tax and consumption taxes. But the incidence of the corporation income tax, the property tax, and payroll taxes is still in considerable dispute. It can be shown that, under perfect competition, price flexibility, perfect factor mobility and inelasticity of the supplies of labor and capital, these three taxes are borne by the factors of the production. This means that the corporation income tax and the general property tax reduce the return on capital in general,[6] and the payroll tax is borne entirely by wage earners. On the other hand, to the extent that business firms and owners of real estate do not operate to maximize profits and have the power to set prices, the corporation income tax, the property tax, and the payroll tax may be shifted in whole or in part to consumers.

---

[2] This assumes that the use of indirect taxes does not alter the distribution of factor incomes.

[3] Nontax receipts include personal and business nontax receipts, some items that are regarded as social insurance receipts in the national income accounts, and an assortment of state-local fees and licenses.

[4] The net receipts of government enterprise were also excluded from the tax concept used in the study, since these receipts are clearly payments for services rendered.

[5] Relative tax burdens could also be measured by comparing taxes with family wealth or net worth. This comparison was not made because income is generally accepted as an appropriate measure of tax-paying ability, and up-to-date and reliable wealth data for the household population are not available.

[6] However, most economists believe that the property tax on raw land is capitalized and falls on those who own the land when the tax is imposed or increased.

TABLE 1—MAJOR INCIDENCE ASSUMPTIONS USED IN THIS STUDY

	Variant I			Variant II		Variant III		
Tax and Method of Allocation	A	B	C[a]	A	B	A	B[b]	C
Individual income tax								
To taxpayers	X	X	X	X	X	X	X	X
Sales and excise taxes								
To consumption of taxed commodities	X	X	X	X	X	X	X	X
Corporation income tax								
To dividend recipients				X	X			
To property income in general	X	X						
Half to dividends; half to property income in general			X					X
Half to dividends; a quarter to consumption; a quarter to employee compensation						X		
Half to property in general; half to consumption							X	
Property tax on land								
To landowners	X			X	X	X	X	X
To property income in general		X	X					
Property tax on improvements								
To shelter and consumption				X	X	X	X	
To property income in general	X	X	X					
Half to shelter and consumption; half to property income in general								X
Payroll tax on employees								
To employee compensation	X	X	X	X	X	X	X	X
Payroll tax on employers								
To employee compensation	X	X	X	X		X		X
Half to employee compensation; half to consumption					X		X	

[a] Most progressive variant.
[b] Least progressive variant.

We did not attempt to settle these issues in the study, but we do provide information on the effects of a sufficient number of different options to permit others to develop estimates of tax burdens on the basis of the particular assumptions they prefer.

Relationship Between Income and Taxes. Although it is not immediately obvious, the total amount of income as well as the taxes paid by each family unit depends on the incidence of the various taxes. Thus, if a tax is assumed to be borne by consumers rather than by labor or capital, or vice versa, family income must be changed. The changes are required for two reasons: first, factor incomes are overstated to the extent that the tax is shifted to consumers; and, second, where taxes are borne by property owners, the amount of taxes allocated to the household sector will depend upon the proportion of total property income received by that sector.[7] Since all the incidence assumptions in the study treat the individual income tax and sales and excise taxes in the same way, modifications in family income were required to maintain consistency between the definitions of income and taxes only for different assumptions regarding the incidence of the corporation income tax, the property tax, and the payroll tax.

---

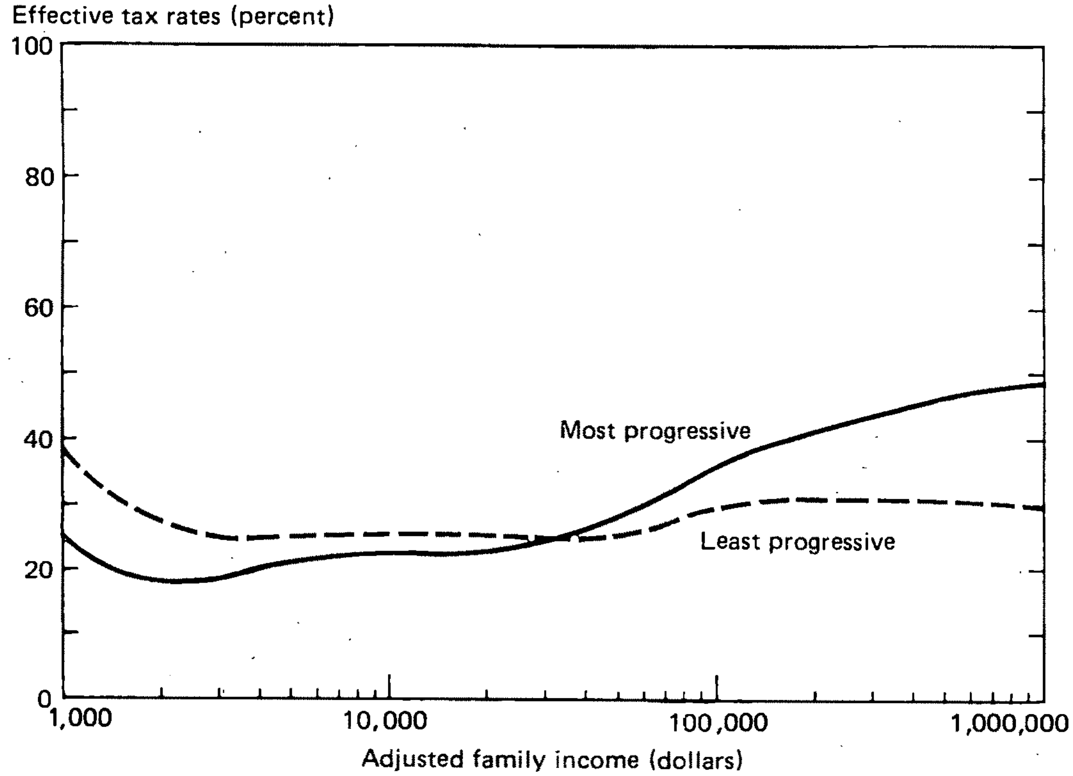
[7] In making these adjustments, it was assumed that consumption patterns and factor shares did not change under the different incidence assumptions.

II. Major Findings on the Distribution of Tax Burdens

The distribution of tax burdens is usually presented by absolute income levels, an arbitrary classification which tends to exaggerate the significance of the data in some parts of the income distribution and to reduce it in other parts. The MERGE file permitted us to prepare tax burden calculations by percentile groups as well as by absolute income levels, and thus to compare tax burdens for groups representing the same number (or percentage) of family units in different parts of the income scale. The pictures of relative tax burdens that emerge from the two classifications are quite different.

Effective Tax Rates. The major conclusions on tax burdens are summarized in Figures 1 and 2. Each shows the effective rates of federal, state, and local tax throughout the scale of incomes for the most progressive and least progressive sets of incidence assumptions used in the study. Effective rates are plotted on a logarithmic scale by absolute income levels in Figure 1 and on an arithmetic scale by income percentiles in Figure 2.

Figure 1 gives the impression that there are large variations in relative tax burdens between low- and middle-income family units and between middle- and high-in-

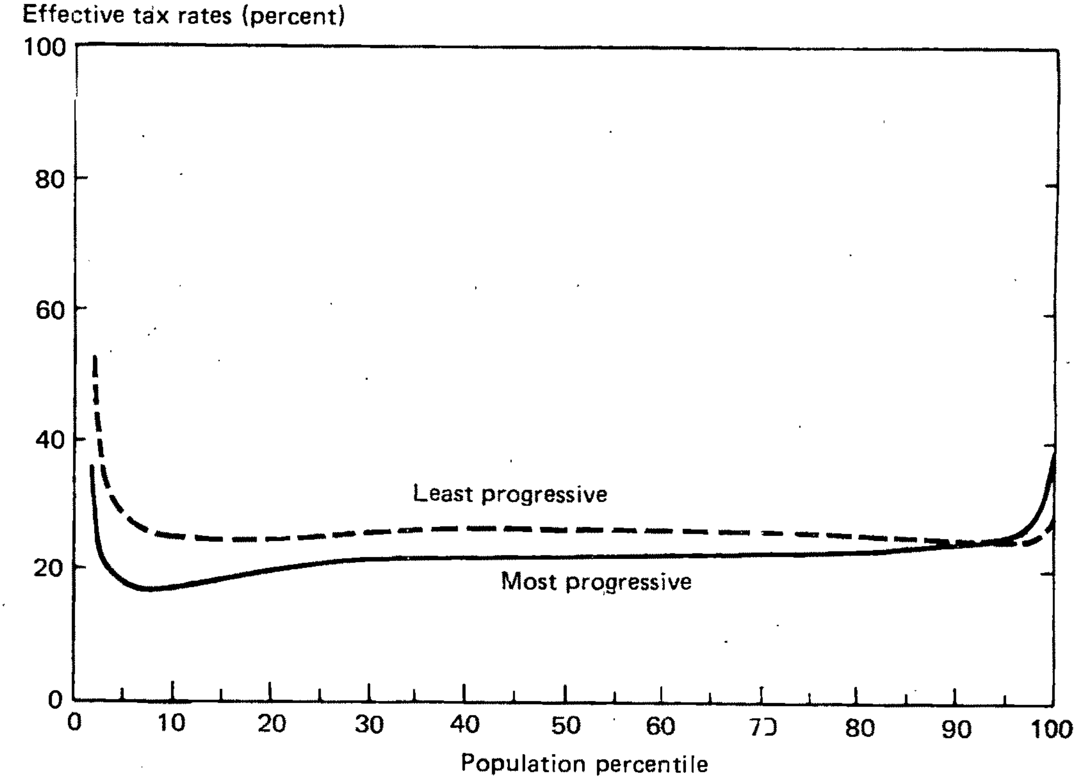


Source: Computed from the 1966 MERGE data file.

FIGURE 1. EFFECTIVE RATES OF FEDERAL, STATE, AND LOCAL TAXES UNDER THE MOST AND LEAST PROGRESSIVE INCIDENCE VARIANTS, BY ADJUSTED FAMILY INCOME CLASSES, 1966

come ones. Under both sets of incidence assumptions, effective tax rates are high at the bottom of the income scale and they then drop until the $3,000 level is reached. Between $3,000 and $25,000, the effective rates range between 20 and 25 percent of income, and they then diverge above the $25,000 level. Under the most progressive set of assumptions, the tax burden rises sharply until it reaches almost 50 percent of income for families with incomes of $1,000,000 or more. Under the least progressive set of assumptions, the tax burden reaches a maximum of only about 30 percent for those with incomes of $100,000 or more.

The effective tax rates under the same two incidence variants are shown in Figure 2 for families classified by percentiles in the income distribution. As can be seen, the picture in Figure 2 is very different from that shown in Figure 1. The difference in effective tax rates under the two variants is very small over practically the entire income scale. Under each variant, there is very little difference in overall effective tax rates between the tenth and ninety-seventh percentiles of family units. For this broad range of the income distribution, which covers incomes between $2,000 and $30,000 and includes 82 percent of all family units, the tax system is either proportional or slightly progressive. At both ends of the distribution, the effective rates rise sharply, but the rise at the top is much more moderate under the least pro-



Source: Computed from the 1966 MERGE data file.

FIGURE 2. EFFECTIVE RATES OF FEDERAL, STATE, AND LOCAL TAXES UNDER THE MOST AND LEAST PROGRESSIVE INCIDENCE VARIANTS, BY POPULATION PERCENTILES, 1966

gressive variant than under the most progressive one.

In addition to differences that arise because of differences in incidence assumptions, there were substantial variations in tax rates among various economic and demographic groups in the population that are due to the structural features of the *U.S.* tax system. For example, homeowners pay lower taxes than renters, urban residents pay somewhat higher taxes than residents of rural-farm areas, and married couples pay lower taxes than single persons.

Perhaps the most interesting calculation we were able to make on the basis of the MERGE file was an estimate of the relative tax burdens imposed on income from labor and capital. As might be expected, income from capital bears a much heavier tax burden than income from labor if the corporation income and property taxes are assumed to be taxes on capital. Under these assumptions, the average tax rate was 33.0 percent on income from capital compared with 17.6 percent on income from labor. But the difference is narrowed considerably if the corporation income and property taxes are assumed to be paid in whole or in part by consumers. In these circumstances, income from capital bears an average tax rate of only 21 0 percent, while labor income bears a tax of 16.0 percent.[8]

Taxes and the Distribution of Income. Because there is so little difference in effective rates for most of the population, the tax system has very little effect on the relative distribution of income in the United States. As expected, the Lorenz curve for the distribution of after-tax income under the most progressive incidence assumptions lies closer to the line of equal distribution than the before-tax curve. But the movement toward equality is relatively small: only 5 percent under the most progressive set of incidence assumptions (and much smaller percentages under any other variants examined in the study). In the case of the least progressive variant, the after-tax distribution of income is only one-quarter of 1 percent more equal than the before-tax distribution.[9]

III. Conclusions

The tax system is virtually proportional for the vast majority of families in the United States. But the nature of incidence assumptions used are crucial in evaluating the progressivity of the overall tax system and its effect on the distribution of income.

Effective tax rates are high both at the bottom and very top of the income scale, but the high rates for those in the lowest income classes are probably not representative of the tax burdens they pay over longer periods, because those classes contain a large proportion of retired persons and individuals with temporarily low incomes.

If it is assumed that the corporation income and property taxes fall on income from capital, the very rich pay high tax rates because a substantial proportion of their income comes from property income. Under this assumption, the tax burden of those with incomes of $1,000,000 is roughly double that paid by most families. If corporation income and property taxes are assumed to be shifted in whole or in part to consumers, the tax burden at the highest income level is about 30 percent, or only 5 percentage points more than the average effective rate paid by most families.

---

[8] The tax rates on both income from capital and labor fall under the least progressive variant because the rate on consumption more than doubles.

[9] The figures refer to percentage changes computed with respect to the before-tax area of inequality. Thus, if the area between the before-tax Lorenz curve and the line of perfect equality is A and the area between the after-tax Lorenz curve and line of perfect equality is B, the percentage change in equality is equal to $(A-B)/A$.

All told, taxes have only a minor impact on the distribution of income in the United States. Even the most progressive set of incidence assumptions produces a pattern of tax burdens that makes the relative distribution of income only 5 percent more equal.

REFERENCE

B. A. Okner, "Constructing a New Data Base from Existing Microdata Sets: the 1966 MERGE File," *Annals of Econ. and Social Measure.*, July 1972, *1*, 325–42.

Income Conditioning in a System of Transfer Programs

By Barry L. Friedman and Leonard J. Hausman*

What is the optimal mix of transfer programs? This question has recently received growing attention. We contribute to the discussion by considering the neglected role of income conditioning in determining the optimum and the view that income subsidies are superior to commodity subsidies. In particular, we demonstrate that in the presence of income conditioning, a small shift from an income subsidy to a commodity subsidy will in some circumstances actually increase both work effort and utility. It should be emphasized at the outset that we do not prove the general superiority of commodity subsidies, but we do prove that income subsidies are not universally superior. Thus the nature of the optimal mix is not obvious, and the optimal program might even include both income and commodity subsidies. In view of this, we conclude with a discussion of some of the practical problems in coordinating a system of more than one transfer program.

I. The Likelihood of Income Conditioning in Transfer Programs

A subsidy is income conditioned if the amount of the subsidy diminishes with rising income. Most of the commonly discussed income subsidies are income conditioned, and some forms of commodity subsidies are too. Consider, for example, a negative income tax (*NIT*) with a positive guarantee level $\overline{Y}$, a nonnegative tax rate t, and a nonnegative benefit level $N=\overline{Y}-tY$, where Y is the nonsubsidy income of the eligible family. Certainly the *NIT*—i.e., N—is income conditioned provided $t>0$; so are other income subsidies such as the demogrant (financed by an income tax) and the credit income tax. The only income subsidy not income conditioned is the lump sum subsidy. Since it is paid equally to all members of a demographic group, it does not concentrate limited resources on the poor and is thus hardly efficient as a tool for reducing poverty.[1]

Several kinds of commodity subsidies need not be income conditioned, however. For example, consider the wage subsidy paid on each unit of the commodity, labor. In order to concentrate benefits among low wage workers, one requires that the wage subsidy paid per unit of labor be $\bar{p}-rp$, where $\bar{p}$ can be interpreted as a minimum guaranteed wage, p is the actual gross wage of the worker, and r is a tax rate. In other words the usual wage subsidy is typically wage conditioned with payments declining and finally ending as a result of rising p. Although income and the wage rate are related, a wage subsidy is not in the strict sense income conditioned because a person can increase his income without changing his subsidy per unit of labor (by working more). One could imagine an analogous subsidy on any com-

---

* Assistant professors, Brandeis University. The work described was financed by a research grant from the Manpower Administration in the U.S. Department of Labor. The views indicated here are those of the authors and not necessarily those of the Manpower Administration.

[1] It is possible to construct a negative income tax with benefits for those in the lowest income range not income conditioned, by making $t(Y)=0$ for all Y below some level Y_s, but positive for higher Y. The minimum taxable income, Y_s, is ordinarily called a set-aside.

modity, expressable by the same formula in terms of the commodity price p. In the case of a housing subsidy, for example, payments could be rent conditioned. Since poor people tend to pay relatively low rents, such a program might be antipoverty efficient. Although income and rent may be related, income may rise without a family choosing to pay more rent so that such a program would not be income conditioned. Indeed, a subsidy conditioned by a commodity's own price will be antipoverty efficient only if poor people always pay (or receive) a lower price than the nonpoor. More generally, some device is needed to concentrate the benefits of a commodity subsidy on the poor, and income may often be the most convenient conditioning variable. We may conclude that although income conditioning is not a necessary element of a subsidy program, it is likely to be part of income subsidy programs and will also appear frequently in antipoverty efficient commodity subsidies.

II. Income Conditioning and the Optimal Mix of Transfer Programs

We shall now consider the consequences of income conditioning for the commodity vs. income subsidy controversy, and present the conclusions of arguments developed elsewhere, but not yet published. (An appendix containing them is available from us on request.) We concentrate on the behavior of a single individual, focusing on work hours, H, and utility, U. Consider an individual eligible for negative income tax benefits whose only income is labor income earned at the prevailing wage rate which for convenience we set equal to one. His benefits are then $N=\overline{Y}-tH$. His total income, $H+N$, can be spent on the two goods, x_1 and x_2, at fixed market prices, p_1 and p_2. Now suppose small subsidies are introduced on each unit of one or both of the commodities purchased. Simultaneously, reduce the benefits under the negative income tax so that the total subsidy the family receives, commodity subsidies plus *NIT* (allowing for a possible change in H), is unchanged. The critical point in the argument is that t, now the cumulative tax rate, is the same before as after the introduction of the commodity subsidies.

Given these assumptions, we derive the change in work effort:[2]

$$(1)\qquad \frac{dH}{H}=\frac{\eta_{H1}\dfrac{dp_1}{p_1}+\eta_{H2}\dfrac{dp_2}{p_2}}{1-t\dfrac{\partial H}{\partial \overline{Y}}}$$

η_{H1} and η_{H2} are the pure substitution effects of p_1 and p_2, respectively, on H, expressed in elasticity form. Since a commodity subsidy is equivalent to a cut in price,

$$-\frac{dp_1}{p_1}\quad\text{and}\quad-\frac{dp_2}{p_2}$$

measure the respective unit subsidies on the two commodities expressed as percentages of price. Since the commodity subsidy payments are offset by the cut in the *NIT*, one might initially expect that the income effects of the two changes might cancel. Indeed, the cut in the guarantee $\overline{Y}$ of the *NIT* has no substitution effect, and we find that the percentage change in H depends on a numerator involving just the two weighted pure substitution effects of the commodity subsidies, η_{H1} and η_{H2}. There is an income effect, however, accounted for by the denominator. The fall in $\overline{Y}$ is less than the amount of the commodity subsidies to the extent that H and thus tH rise. But then the work discouragement from the income effect of the commodity subsidies outweighs the encouragement from the reduced $\bar{Y}$, making dH/H smaller than the numerator alone would predict. If leisure is a normal good, the income effect on H will be negative, so the denominator will exceed one when $t>0$.

---

[2] Although our problem is different, we use a method similar to that of W. J. Cortlett and D. C. Hague in the direct vs. indirect tax controversy.

Consider now whether a shift to one or more commodity subsidies can increase work effort. The denominator of (1) is always positive. Since subsidies mean negative dp_1 and dp_2, η_{H1} must be negative if a subsidy on commodity 1 is to increase H, and similarly for commodity 2. It is known that all the pure substitution effects for a single commodity, expressed in elasticity form, must add to zero:

$$\eta_{H1} + \eta_{H2} + \eta_{HH} = 0 \tag{2}$$

where η_{HH} is the pure substitution effect of wage rate on H. But since η_{HH} is necessarily positive, $(\eta_{H1}+\eta_{H2})$ must be negative. Thus at least one of the two must be negative, so that at least one good must be rival with leisure (or complementary with work effort). Indeed, the larger η_{HH} the greater the likelihood that at least one of the η_{Hi} $(i=1$ or $2)$ will be large. It follows that there is always some commodity subsidy which will tend to increase work effort. Notice that this result holds even when $t=0$, so that commodity subsidies can increase work effort even in the absence of income conditioning.

We are not aware of empirical estimates of the cross elasticities η_{Hi} for any good i so we cannot single out particular commodity subsidies that would most strongly affect work effort. Nevertheless, some information can be derived from η_{HH} itself. Orley Ashenfelter's rough estimates suggest that, for married women, $\eta_{HH}=.8$ and $\partial H/\partial \bar{Y} = -.10$. Assume that all commodity prices are cut by 1 percent (a 1 percent subsidy). Then if the tax rate is .66, formula (1) tells us that H will increase by .75 percent. If $\eta_{Hi}>0$ for any good, the effect on H can be increased by dropping the subsidy on that good. While not a strong result, this raises at least a possibility that some commodity subsidies might have a nonnegligible effect on work effort. For males, Ashenfelter's reasonable estimates are $\eta_{HH}=.15$ and $\partial H/\partial \bar{Y}=-.10$. The 1 percent uniform subsidy would then increase work effort by .14 percent at a tax rate of .66.

When we consider the effect of the shift to commodity subsidies on utility, income conditioning becomes critical. Our unpublished arguments lead to the result that

$$dU = \lambda t dH \tag{3}$$

where λ may be interpreted as the marginal utility of income, and dH is the change in H resulting from the particular trade-off of programs we have assumed. As long as $t>0$, any such shift that increases H will increase utility. But as we have just demonstrated, there are likely to be commodity subsidies that will increase H. The shift from an *NIT* to such subsidies will thus increase both H and utility. The effect on utility will be stronger the larger t. When $t=0$, the shift leaves utility unchanged no matter what happens to H. This is somewhat surprising since the usual demonstration (see, e.g., Henry J. Aaron and G. M. von Furstenberg) that an income subsidy is superior to a commodity subsidy supposedly holds for the case $t=0$. In fact dU is the first order term in the Taylor expansion of ΔU. It is true that $dU=0$ for $t=0$, and if only very small changes are considered, higher order terms can be ignored. But as Harold Hotelling demonstrated, the superiority of the lump sum subsidy or tax depends on the second and possibly higher order terms of the expansion. We have demonstrated that when $t>0$, the first order term dU may actually be positive. When the shift in programs is very small this term predominates while for a larger shift it will tend to counteract the opposite second order effect. We have thus established the possibility that the shift to commodity subsidies might increase utility in the presence of income conditioning.

To understand why H affects dU, consider total income, $C+\bar{Y}+(1-t)H$, where C is the amount of the commodity subsidies. Ordinarily, an increase in after-tax employment income, $(1-t)H$, is achieved at the expense of leisure and cannot be expected to increase utility. On the other hand, consider nonemployment income, $C+\bar{Y}$. If the cut in the *NIT* benefit $(\bar{Y}-tH)$, must equal C, then $C+\bar{Y}$ will not remain constant if H increases, but will increase by tdH. Multiplying by λ converts the money sum tdH into units of utility. The increment in utility thus results from an increase in nonemployment income which, given the income conditioning of our programs, happens when work effort is increased.

We can thus prove two points concerning a shift to a small commodity subsidy and an offsetting cut in an *NIT*. First, it will ordinarily be possible to find commodities to subsidize such that work effort will rise. That is, there exist commodity subsidies that will have a stronger effect on work effort than an equal *NIT*. This result does not depend on income conditioning. Second, if the shift does succeed in increasing work effort, then in the presence of income conditioning it will increase utility, at least for a sufficiently small shift. We have not proved the superiority of one transfer system over another. Rather, we have demonstrated that even the response of a single individual to alternative programs cannot be predicted by theoretical argument alone but must be estimated empirically from information on how commodity prices affect work effort. Indeed, information concerning individual response would then have to fit into a broader model allowing for (possibly differing) administrative costs and including a social welfare function before any conclusions could be reached on the optimal mix of transfer programs.

III. Problems in the Coordination of Multiple Transfer Programs

Political considerations as well as our arguments of the last section might call for a system of more than one transfer program. In this section we will consider some of the problems that will arise when more than one program is desired. First, we briefly review the cumulative tax rate problem resulting from two or more programs being income conditioned. Second, we will inspect the joint accounting system and some of its consequences.

The cumulative tax problem can be understood by considering an income subsidy and a commodity subsidy, both income conditioned. It can be demonstrated that the cumulative tax rate on income that results from the two programs is the sum of their tax rates. Likewise, the cumulative guarantee is the simple sum of the two guarantees. (See Thad W. Mirer.) It is possible to integrate programs so as to reduce the cumulative tax rate somewhat while keeping the individual program guarantees and tax rates fixed; but this would also entail a reduction in the cumulative guarantee. An alternative program combination could use a wage subsidy to offset part of the tax rate of a program (or programs) while leaving the guarantee unchanged.

To develop a sense of the practical constraints facing designers of a new mix of transfers, we present data on existing cumulative guarantees, tax rates, and breakeven levels. By 1973, virtually all poor female-headed households received Aid to Families with Dependent Children (*AFDC*) payments throughout the year. (See Barbara Boland.) *AFDC* payments also were made, at any point in time, to roughly one-fourth of male-headed families in poverty. *AFDC* families generally are entitled to Medicaid and Food Stamp benefits. Only crude estimates can be made of the average cumulative guarantee from these three programs. Data on average net benefits per recipient suggests that, for a family of four, the combined *AFDC*/Food Stamp annual guarantee exceeds $3,500,[3] while average Medicaid benefits exceed $500. Calculations are complicated further by the liberal tax credit that *AFDC* recipients are given for work expenses. *AFDC* benefits for female-headed families are reduced at an average rate of perhaps 40 percent of gross earnings. Nationwide, breakeven levels of earnings for the latter group may average more than $6,500 in *AFDC* and somewhat less in Food Stamps. Moreover, use of data on average guarantees and tax rates conceals important problems raised by atypical, but numerous, cases. In New York, for example, the *AFDC*/Food Stamp guarantee is $4,300; average Medicaid aid is $1,200. In Missouri the annual *AFDC* guarantee is only $1,560. But average tax rates are so low that the breakeven level exceeds $8,600. The New York *AFDC* breakeven level, because of its high guarantee, is roughly $10,000. An *NIT* with a guarantee of $3,500 and a breakeven level of income of $6,500 would have a 54 percent average tax rate—and still would reduce current net benefits for large groups of families. In 1974, with a social security tax rate in excess of 5 percent and a federal income tax rate of 14 percent starting at $4,300 of annual income, the average cumulative tax rate on earnings never would be less than 60 percent and frequently would approach 75 percent.[4]

An additional significant problem in program design is the income accounting system which refers to the frequency with which income is reported and benefits correspondingly adjusted. We know that a program with monthly accounting and benefit adjustment provides a recipient with an incentive to bunch his earnings into a portion of the year so that he can receive maximum benefits during the rest of the year. In a multiprogram system, it is conceivable that different programs will have different accounting systems. The program with short accounting and memory periods will be more responsive, but the one with longer periods will be more equitable. It can be demonstrated that when one program has monthly accounting and another one has annual accounting, the incentive to bunch earnings can be significantly reduced if the benefits under the monthly accounting program are taxed by the annual accounting program. The monthly program is still responsive to short-term needs, but if the needs persist, the tax in the annual program reduces some, but not all, of the undesirable (and costly) incentive to bunch.

This case illustrates how some problems may be eased by a coordinated accounting system. In addition, joint administration of work tests and family unit rules might lead to significant economies in a multiprogram system. Some joint administration already exists for *AFDC* families who also receive Food Stamps and Medicaid.

Given the guarantee levels that low income groups have come to expect, low tax rates will extend benefits to very large groups of people. If the tax rate is too high, it may not only jeopardize work effort, but also eliminate net benefits to current recipients. Given prevailing standards, options are few and costly. One response to this has been a proposal for categorical programs. Employables would be covered either under a wage subsidy or under public service employment. Those classified as

---

[3] This results from an average *AFDC* guarantee of $2,600 and Food Stamp guarantee of $1,700. They are integrated via the "sequencing" approach, under which the latter program includes *AFDC* benefits in its definition of taxable income.

[4] The indicated guarantees or tax rates ignore other programs, like public housing and day care. These offer high benefits and must be income conditioned if costs are to be kept within bounds.

THE GREAT SOCIETY VERSUS THE NEW FEDERALISM

Sorting Out the Social Grant Programs: An Economist's Criteria

By CHARLES L. SCHULTZE*

Expenditure for federal categorical grants for social programs grew substantially in the 1960's: from $1.3 billion in fiscal year (*FY*) 1960 to $8.8 billion in *FY* 1970 to $14.9 billion in *FY* 1973 (see Edward R. Fried et al., p. 5, Table 1-1). Some argue that such grants are competitive with revenue sharing as a means of affecting the production and distribution of collective goods at the state and local level.[1] Indeed, the Nixon Administration has spent much effort seeking to consolidate many of these grants into broad bloc grants, called "special revenue sharing," arguing that the proliferation of categorical grants has produced few substantive results, entails an unnecessary degree of federal control over the use of grant funds and has created a jungle of red tape and confusion.

Without denying that the current grant system is often cumbersome and sometimes inefficient, proponents fear that its replacement by broad bloc grants would result in a substantially less egalitarian distribution of public services and render impossible the provision of some important types of collective goods.

Two different theoretical issues are mixed up in this debate: First, to what extent does a specific federal intervention into the quantity or quality of goods and services produced at the state and local level address a legitimate need, either to achieve some desired distribution of income or to compensate for some form of market failure?[2] Second, what combination of policy instruments is most efficient for the purpose at hand—i.e., what are the criteria for using categorical grants-in-aid as opposed to other techniques of intervention?[3]

I. Categorical Grants and the Theory of Fiscal Federalism

A theoretical analysis of the proper role of categorical social grants might start from (but not, as I shall argue presently, end up with) the literature on the theory of fiscal federalism.[4] One branch of this theory concerns the optimal layering and appropriate jurisdictions of a multilevel set of governments engaged in the produc-

---

* Brookings Institution and University of Maryland. I thank my colleagues Alice Rivlin, Ned Gramlich, Robert Hartman, Henry Aaron and Martin McGuire for their help and suggestions.

[1] For a discussion of transfers-in-kind and cash transfers, which may be viewed as alternative means of redistributing income among individuals, see Lester Thurow. Expenditures for transfers-in-kind have also grown considerably, from $0.7 billion in *FY* 1960 to $13.6 billion in *FY* 1970 to $21.7 billion in *FY* 1973 (see Fried et al., p. 5, Table 1-1; figures exclude veterans education benefits).

[2] The term "market failure" includes cases of less than optimal allocation on the part of state and local governments.

[3] I do not attempt to evaluate specific programs, except in the context of these two criteria.

[4] For a convenient summary of the current state of the theory and a useful set of references, see Wallace E. Oates.

tion of public goods, focusing on the nonappropriability characteristics of the benefits from public goods and the Samuelsonian equation of marginal costs with the sum of marginal rates of substitution. In the case of a pure public good, extension of jurisdictional boundaries decreases the per capita costs of provision, but leads to welfare losses if the extension involves population groups with systematically differing preferences for the jointly supplied good. More importantly, for goods which are "impurely" public, such that extension of the population covered requires additional costs to prevent "congestion" or to maintain the quality of the service provided, the welfare loss from extension is even larger. The optimal jurisdictional arrangement depends upon the shape of the functions describing these offsetting welfare gains and losses. And since the functions are not the same for all public goods, the optimal jurisdictional boundaries may be different for different public goods, but the optimal diversity is limited by the costs of voting and decision making, the high overhead costs associated with many geographically overlapping jurisdictions, and spillover effects among public goods.

The second branch of the theory of fiscal federalism, which considers the circumstances that require intergovernmental fiscal transfers, calls for two kinds of grants-in-aid from higher to lower levels of government. Unconditional transfers of fiscal aid ("revenue sharing") may be necessary to deal with the equity and efficiency consequences of differences among jurisdictions in income and wealth;[5] categorical aids, in which the granting government specifies the use of the funds, are justified in the theory as Pigovian subsidies designed to deal with jurisdictional benefit spillovers. The Pigovian basis for categorical grants generally implies the granting of an open-ended matching subsidy, equivalent to a price reduction, designed to induce state or local governments to increase the provision of the particular public good. There is some controversy over the extent to which spillovers lead to potential "gains-in-trade" among local jurisdictions, that can be realized by bargaining and negotiation without grants from higher level governments.[6] Where spillovers affect only a small and geographically adjacent set of jurisdictions, planning grants rather than Pigovian subsidies may be called for. In general, however, there is a wide consensus that spillovers which affect many jurisdictions—e.g., the external benefits of education in a very mobile society—cannot be handled by bargaining because of the very large costs involved and the free rider problem, and therefore require subsidies.

In fact most categorical grants for social programs are not based on this rationale, and whatever the rationale, virtually none takes the form of Pigovian subsidies. Indeed, the only significant open-ended matching grants are public assistance (a cash transfer) and Medicaid (a transfer-in-kind); neither is a subsidy designed to deal with public goods spillovers. On the other hand, a large proportion of federal social grants provide a specific sum of money, with a very high federal matching ratio (80, 90, or 100 percent) for a very particular purpose, and specify the delivery of very particular forms of collective goods which had not been previously furnished by grant recipients. Many are subject to guidelines, restrictions, and other requirements dictating in great detail the nature of the good or service to be provided.

Thus, the paradigm of grants as price subsidies to an independent decision-making unit designed to change the incentives it faces is not very useful for analyzing most of the existing social grants. Rather, many of these grants are a means by which the federal government uses state and local governments (or in some cases local nonprofit organizations) as agents or subcontractors to produce centrally determined amounts and kinds of collective goods,[7] since, for a number of reasons, principally historical and political, the federal government itself virtually never delivers collective goods or services at the local level. In fact, the contracting paradigm comes closer to the real situation. It is to the theory of public finance in general, rather than to the specific theory of fiscal federalism, that we must turn for an analysis of the categorical grant system.

---

[5] See Richard Musgrave (1961a, 1961b) and James Buchanan.

[6] See Ronald Coase and Buchanan and William Stubblebine.

II. Objectives of the Social Grant Programs

The $15 billion of the federal social grant program appear to serve seven major kinds of objectives:

1. *Transfers-in-kind of collective goods to low income or other target population groups.* Two kinds of collective goods fall into this category: first, public goods (in the theoretical sense of the term "public") such as social services for welfare recipients, which by their nature are not practically marketable; and second, private goods produced collectively and distributed free, such as the services of neighborhood health centers, community mental health centers or manpower training programs.

Just as the private market generates a distribution of money incomes which public policy seeks to correct, sometimes through transfers-in-kind of market goods, so does the political mechanism of state and local government generate a distribution of collective goods which the federal government seeks to correct, sometimes through grants. It is not my aim here to provide a rounded theoretical underpinning for the widely held view among federal legislators that state and local governments undersupply collective goods consumed principally by low-income residents. But it is worth noting that local taxpayers do have a self-interest in increasing the proportion of high-income residents in their community, and taxes levied to support collective goods which principally benefit low income recipients put them at a competitive disadvantage. Even if state and local public goods were typically financed by benefit taxation, a concentration of high-income residents in a community would tend to benefit each taxpayer (so long as the income elasticity of demand for most public goods is positive). A voter might be perfectly willing to incur the added taxes to provide collective goods for the disadvantaged. But to the extent that the higher tax rates threatened to reduce the ratio of high- to low-income residents in the community, this redistributive decision would cost the voter still further tax increases. Hence it is perfectly rational for a citizen to vote against such programs for his own community and yet vote for the provision of the same program by the central government. Even if no jurisdictional spillovers exist, therefore, the income distribution goals of society cannot be met without the central provision of certain collective goods.

2. *Spreading innovations in the delivery of public services.* As with private production, research and development and experimentation with potential improvements in the production of locally delivered collective goods may entail externalities. Moreover, the market for many private goods is national, served by national firms, and innovations there, once proven, tend to spread quickly. This is not the case with local collective goods,

---

[7] By collective goods I mean those goods actually produced and distributed free (or at highly subsidized prices) by governmental organizations. In terms of appropriability and excludability some of them may be "public" and some "private."

since markets for them are neither competitive nor national. Such considerations provide a rationale for federal grants for community mental health centers, neighborhood health centers, and even private health maintenance organizations.

3. *Speeding labor force supply responses.* The creation or rapid expansion, under federal sponsorship, of drug abuse and alcoholism centers, family planning clinics, community mental health centers, and similar institutions has sharply increased the demand for a wide variety of professional and semiprofessional skills. Presumably the rise in demand would ultimately call forth a supply response. But a deliberate stimulation of that response can increase the efficiency with which the services are delivered and reduce the rents paid out in the interim. The federal government, therefore, accompanies many of its service grant programs with training grants for specific kinds of skills.

4. *Improving the flow of information and influencing the preferences of some groups of consumers.* While family planning clinics subsidize the use of contraceptive devices for low-income groups, their major aim is to provide information and indeed to proselytize, especially in low-income communities. Neighborhood service centers which provide information about the availability of social services and alcohol and drug abuse programs share some of these same characteristics. Governmental provision of information, which improves the efficiency of consumer choice, is a standard means of improving market performance. The deliberate attempt to change tastes is more difficult to rationalize in a theoretical framework grounded on the sanctity of consumer preferences. The autonomy and supremacy of individual preferences, however, are concepts which appear to inhibit no one except economic welfare theorists. Changing those preferences is an important component of many public policies and forms a significant part of the educational content of the public (and private) school system.

If one grants the logic of the rationale, most of the grants which pursue this objective also fall within the scope of category (1) above—collective goods particularly addressed to disadvantaged groups, which will generally be undersupplied by local governments.

5. *Increasing the bargaining power of the poor and minorities at local level.* Local decisions about the provision of collective goods are made not merely on the basis of electoral choices (one man, one vote) but also in response to the influence of organized groups. In elections voters "buy" very comprehensive packages of joint products when they choose a particular candidate. The joint products are partially decoupled through the operation of formal and informal pressure groups, whose influence stems from campaign contributions, "old-boy" relationships, and participation in a log-rolling process. Typically, but not universally, the groups are organized around a producer interest (real estate, retail trade, banking, etc.). The decisions of community officials which affect any organized group are virtually never made without their participation. To the extent that log-rolling, side payments, and group bargaining are a means of aggregating individual preferences in the allocation of resources to collective goods, both efficiency and equity require the representation of all relevant groups in the process. The federally supported Community Action Program was grounded on the judgment that without assistance the poor and especially the minority poor are unlikely to be fairly represented in this scheme of things. Many other grant programs targeted on the poor—e.g., Model Cities, Title I of the Elementary and Secondary Education Act (*ESEA*)—require that representatives of the affected

low-income population participate in decision making.

6. *Overcoming problems raised by jurisdictional boundaries: planning grants.* Where externalities in the provision of collective goods are largely confined to a limited number of adjacent jurisdictions, subsidies to internalize those externalities may not be required. Even with efficient jurisdictional boundaries such externalities are likely to occur: in transportation and land use; in the provision of highly capital-intensive health-care facilities; and in many other areas. Actual jurisdictional boundaries are far from efficient and usually very difficult to change for a variety of political and constitutional reasons. Those same obstacles often raise a barrier to the "spontaneous" development of the planning bodies necessary for dealing with adjacent externality problems. Federal provision of grants to appropriate multijurisdictional planning bodies can be a device to encourage and accelerate the necessary negotiation and bargaining.

7. *Overcoming problems raised by jurisdictional boundaries: Pigovian subsidies.* As I have already pointed out, most of the federal social grants do not easily fall within this category. Most federal grants for elementary and secondary education, for example, do not: the largest such grant, Title I of *ESEA*, is aimed at compensatory education of the disadvantaged and fits within category 1 (although some proponents would claim that there are significant extrajurisdictional benefits from compensatory education). Some of the older disease control grants are addressed to a classical spillover problem, and the basis for federal rather than purely state or local support for training medical personnel is essentially a spillover rationale. Federal manpower training programs heavily emphasize the needs of the disadvantaged worker, but while mobility does create spillovers, dealing with these is not the principal objective of the training grants.

The point of the classification scheme is not to suggest that welfare criteria can justify all of the social grants. Indeed, it can be argued that some federal grants appear principally to reflect the simple substitution of central government preferences for state or local preferences and are analogous to merit wants in the case of private goods. Grants for audiovisual equipment in elementary and secondary schools, vocational education, community library support, and law enforcement assistance are examples of programs which might be classified under this heading. But the range of categorical grant objectives which meet welfare criteria does extend well beyond that offered by the theoretical literature on fiscal federalism and must be examined in the broader context of welfare theory relating to the exercise of collective action.

Except in a few cases, the Administration's general and special revenue-sharing proposals do not represent alternative means of accomplishing more efficiently the same sets of objectives sought under existing categorical grant programs. At first blush it might appear that the various special revenue-sharing proposals would convert the existing narrow categorical grants into a system of broad Pigovian subsidies, one for each major governmental function with spillover effects. But this is not the case. The elementary and secondary education special revenue-sharing proposal maintains very many of the categorical restrictions of the existing system. It is, in effect, a modest and probably desirable reform which still retains many of the features of the current programs. While the other proposals represent a more radical consolidation, they are not price-reducing grants. They require no state and local matching and they are not open-ended. Rather, they distribute a fixed amount of funds by formula. In many instances, it is quite likely that after a transition period, these grants would become fungible resources to most recipients and have effects very much like a series of general revenue-sharing grants, each with a differing distribution formula. General and special revenue sharing therefore should be seen as competing with the existing grant system for scarce budgetary resources in the same sense as health programs compete with transportation and space programs. It is a competition among ends, not among alternative means of achieving the same ends.

III. Systemic Problems in the Categorical Grant System

The system of categorical grants currently employed to meet the various objectives spelled out above is characterized by different types of inefficiencies or inequities:

1. *Confusion of objectives.* Some of the programs combine demand- and supply-oriented intervention, but are not predicated on any real evaluation of the desirability of the merger. For example, in supporting neighborhood health centers in the inner cities, community health centers, and day care programs, the federal government both delivers free services to the poor and seeks to develop new institutions for the delivery of those services. As a consequence, the test of the market cannot be applied to the new institutions stimulated by the grants, and the federal government may find itself permanently involved, not simply in the promotion of innovation, but in the detailed day-to-day production of services on a permanent basis.

2. *Supply side inefficiencies.* In using state and local governments as agents in the delivery of local services, the federal government relies on complex regulations, guidelines, negotiations over numerous individual projects, and, more generally, joint decision making with thousands of state and local officials. Necessarily this leads not only to large administrative costs but also to game strategies, grantsmanship, and other phenomena which cause specific projects to be inefficient in size or design.

3. *Demand side inefficiencies.* As pointed out above, those grants whose rationale is based solely on central government "merit wants" for collective goods reduce Pareto efficiency in the provision of collective goods, almost by definition, if preferences vary significantly from locality to locality.

4. *Inequities of underfunding.* The large proliferation of specific grant programs during the 1960's, in a fairly tight budget, has led to a vast amount of underfunding: the collective services are provided free, but only to a fraction of those eligible for them. Rationing is accomplished not by the systematic establishment of tighter eligibility criteria, but more or less arbitrarily by various administrative and political decisions about the distribution of grants. Neighborhood health centers, day care and Head Start programs, manpower training centers, and the various community development grants of Housing and Urban Development are examples of this phenomenon. Perhaps some underlying political process works to distribute the shortages fairly ("If you don't get a model cities program you do get a health center"), but I have seen no evidence to support this hypothesis.

5. *The special problem of capital grants.* The Congress likes to use capital grants to subsidize the provision of collective services: the urban mass transit program, the Hill-Burton hospital program, construction grants for community mental health centers, etc. Subsidies for the purchase of inputs, rather than the provision of outputs, clearly bias producers towards inefficient combinations of inputs. For example,

a recent study of Chicago and Cleveland bus operations concluded that the capital subsidy induced uneconomic early replacement of buses (instead of expenditures on repairs and maintenance) to a degree which raised the costs of bus operations by 23 percent.[8]

IV. Suggestions for Improving the System

Even its proponents agree that the categorical grant system needs reform. A very stark outline of suggestions for reform would include the following.[9]

First, activities under grant programs concerned with the free collective provision of private goods—e.g., neighborhood health centers, community mental health centers, day care and Head Start programs and many of the manpower training programs—should be examined to determine the extent to which they could be split into two or possibly three components: income-tested vouchers for the particular service; "seed money" grants for the initiation of improved means of delivering services; and, in some cases, loan guarantees to provide social risk pooling for new kinds of nonprofit or quasi-governmental institutions which have restricted access to the capital market.

Second, many of the separate developmental and innovative grants which serve the same purpose should be consolidated into a limited number of flexible instruments. In the health services area, for example, there are currently five major types of innovative or developmental grants, some supporting initial development, some ongoing services, and some capital construction: the neighborhood health and community mental health center grants mentioned above, that part of the Hill-Burton hospital construction program which now concentrates on the construction of outpatient facilities in central cities, rural health care grants, and developmental grants for Health Maintenance Organizations. What is needed is a single broad grant program under which the federal government can furnish seed money for development of more effective means of delivering health care, particularly to the poor, according to the geography and the circumstances of particular areas. With an income-tested health insurance voucher scheme, the institutions developed should ultimately be required to meet the test of the marketplace, financing the cost of care (including capital costs) from charges. (In the case of health programs, the last two suggestions should be applied as part of a reform of the present Medicaid system of providing medical insurance for the poor rather than in isolation.) But with an income-related system of health insurance in effect, both equity and efficiency would seem to call for eliminating the free provision of services by neighborhood health centers and community mental health clinics.

Third, many grants provide public goods for the poor or for special groups, or provide private goods with large external public benefits. Recovering costs through charges is impossible or impractical. Family planning clinics, drug abuse and alcoholism centers, and juvenile delinquency prevention programs are examples. If it is decided that redistributive or efficiency gains are large enough to justify such a program, no alternative may be available but to continue federal grant financing of continuing services.

Fourth, narrow categorical grants for continuing support of state and local collective services which are neither innovative nor redistributive to lower income groups are either responses to jurisdictional spillovers or simply federal "merit wants." I very much doubt that a spillover rationale justifies narrow specification of inputs or outputs; and while the view that local public education has net positive jurisdictional externalities is not without foundation, it is difficult on efficiency grounds to support existing narrow grants for programs such as language training and area studies, university-community services, undergraduate instructional equipment, school library resources or equipment and minor remodeling. Spillover benefits from education can be handled by a broad functional grant. Narrow categorical grants that merely satisfy federal merit wants for local collective goods should be eliminated.

---

[8] See William B. Tye.

[9] Many of the suggestions made in this section are developed more fully and with a wide range of illustrations in Fried et al., pp. 170–232.

Fifth, capital grants should be converted into grants based on units of services rendered, and simultaneously reviewed in the light of the preceding criteria. To the extent that outpatient clinics at city hospitals are seen as deserving of developmental support, that support should not be given through Hill-Burton construction grants but, like other developmental funds, should be based in some fashion on output-related criteria. Although this paper is not principally concerned with federal grants for transportation and environmental control, the construction subsidies for mass transit and for municipal waste treatment plants are highly inefficient; other instruments of federal intervention are needed.[10]

V. Conclusion

In conclusion, let me summarize this discussion not by restating it but by giving a brief illustration and extension of the proposals advanced here in the context of reform of urban development grants. The continuation of a number of such grants—Model Cities, neighborhood facilities, open space purchases, urban renewal, and basic sewer and water facilities—as separate categories may be inefficient. While federal government aid to central cities for rehabilitation and development may perhaps be justified on distributive and equity grounds, there is much less warrant for federal specification of the mix of developmental investments. There is, however, one theoretical rationale for some of the categorical specifications. For the reasons outlined earlier, it is at least possible that in many communities the interests of the poor or minority groups may be given short shrift in community development plans.[11] Two kinds of federal categorical specifications deal with this problem: geographical specification of the use of grant funds in a way which insures that most of them are a net addition to the funds spent in poor neighborhoods (model cities) and other attempts to protect the interests of the poor (relocation of displaced residents, minimum residential requirements for urban renewal plans, etc.). In this kind of situation, reform would eliminate *functional* categorization but preserve or even strengthen certain *targeting* requirements, so that a reasonable proportion of the consolidated grant is used for the benefit of low-income neighborhoods. Given the fungibility properties of federal grants in state and local budgets, it may be impossible to design effective targeting features in consolidated urban development grants. But an evaluation of the possibilities would be worth the analytical resources involved.

Because of the inherent difficulties of pursuing many of the objectives of the social grant programs, even the most successful restructuring would leave a significant and irreducible set of inequities and inefficiencies. This should be no surprise in an "n^{th} best" world and does not itself justify dismantling the affected programs. The instruments for accomplishing one set of ends may necessarily be less efficient in the narrow sense than the instruments available to meet other ends. This is a factor to be taken into account in budget allocation decisions as between revenue sharing and social grants. But it is simply one factor among many and need not be the dominant one.

---

[10] See, for example, Fried et al., pp. 238–257.

[11] See, for example, Anthony Downs.

REFERENCES

J. **Buchanan,** "Comment," in National Bureau of Economic Research, *Public Finance: Needs, Sources and Utilization,* Princeton 1961.

——— and **W. Stubblebine,** "Externality," *Economica,* Nov. 1962, *29,* 371–384.

R. **Coase,** "The Problem of Social Cost," *J. Law and Econ.,* Oct. 1960, *3,* 1–44.

A. **Downs,** "Uncompensated Non-Construction Costs Which Urban Highways Impose Upon Residential Households," in J. Margolis, ed., *The Analysis of Public Output,* New York 1970.

E. R. **Fried** et al., *Setting National Priorities: The 1974 Budget,* Washington 1973.

R. **Musgrave,** "Approaches to Fiscal Theory of Political Federalism," in National Bureau of Economic Research, *Public Finance: Needs, Sources and Utilization,* Princeton 1961a.

———, "Reply," in National Bureau of Economic Research, *Public Finance: Needs, Sources and Utilization,* Princeton 1961b.

W. E. **Oates,** *Fiscal Federalism,* New York 1972.

L. C. **Thurow,** "Cash Versus In-Kind Transfers," *Amer. Econ. Rev.,* May 1974.

W. B. **Tye,** "The Capital Grant as a Subsidy Device: The Case Study of Urban Mass Transportation," in Joint Economic Committee, 93rd Congress, 1st Session, *The Economics of Federal Subsidy Programs,* 1973.

Cash Versus In-Kind Transfers

By LESTER C. THUROW*

Except in cases related to pure public goods (which are hard to find in the domestic area), the received wisdom of economics seems to support transfers in cash (the 'New Federalism') rather than in kind. Governments may modify the market distribution of purchasing power, but then should stand aside and allow consumer sovereignty plus competitive markets to work. Any further intervention lowers consumer utility below what it otherwise could be. In a federal government, presumably an ideal system of cash transfers would consist of unrestricted cash transfers from the federal government to individuals and to state and local governments, who would in turn make unrestricted cash transfers to individuals. The result would be an optimum distribution of income within each area with each individual free to make his own expenditure decisions.

But is this common doctrine correct? If not, where should we draw the line between the use of cash and the restricted transfer of particular goods and services?

I. Pareto Optimality

Most of the existing economics literature analyzes the problem from the point of view of Pareto optimality.[1] The potential donor (taxpayer) and the potential donee (welfare recipient) both have utility functions that they wish to maximize subject to a budget constraint. For transfers to be justified, either the donee's utility or his consumption of some particular good must appear in the utility function of the donor. For utility functions to be interdependent in this way, some external benefit is hypothesized to flow from the donee's utility or consumption to the donor.

The state (rather than private charity) must be used to internalize the hypothesized externality since the external benefits are thought to have the nonexcludability and nonrivalability characteristics of public goods.[2] The individual demand curves must be added vertically to determine optimum transfers, and the political process must overcome the revealed preference problem.

From the donee's perspective, unrestricted cash transfers maximize utility. Maximizing the donor's utility is slightly more complicated. If the donee's utility appears in the donor's utility function, then cash transfers also maximize the donor's utility. If the externality flows from some particular good, then the donor may maximize his utility per dollar of transfer either by giving cash transfers and taking advantage of the donee's income elasticity of demand for the targeted good, or by giving a restricted grant (either in the form of vouchers or subsidized in-kind provision)[3] that takes advantage of the donee's price elasticity of demand for the targeted good.[4] Since operating via the price elasticity of demand will always be cheaper than operating via the income elasticity of demand, *restricted transfers always dominate unrestricted transfers when*

---

* Massachusetts Institute of Technology.

[1] For two of many examples see M. Pauly and Harold M. Hochman and James O. Rodgers.

[2] Externalities do not make a good into a pure public good that must be distributed by government. All, or most, domestic public goods could technically be sold privately.

[3] The two are economically identical although they differ in who carries the administrative burden and the ultimate ownership of the housing in question.

[4] For an optimum system of subsidies using both price and income elasticities of demand see Lester C. Thurow (1966).

the interdependence flows from a particular good.

Thus externalities, or interdependencies, appear to account for the provision of restricted transfers, or do they? The Pareto optimality literature argues—sometimes explicitly, and even more often implicitly—that a "rational" donor should only be interested in the donee's utility,[5] and that to want to specify the consumption of particular goods is to be "irrational" and in the world of the "second best." While logically there is no such thing as an irrational preference from the point of view of Pareto optimality, this argument is understandable if the term "externality" is being used as it often is as a vague label for some unspecified psychic income received by the donor. For, so defined, psychic externalities do not really explain why individuals are interested in the education, housing and medical consumption of other individuals. They are merely another way of describing an apparent interest. To be a useful explanation of why utility functions are interdependent, the concept must be limited to physical externalities. These may be sociological (less crime, etc.) or political (better voters, etc.) but they must be real, measurable phenomena.

Even the concept of physical externalities, however, is simply not a convincing explanation of the provision of most domestic public goods. Once a society gets beyond basic public health measures and communicable diseases, medical care does not generate externalities. Death is the most private of all activities, and an individual's health has no nonmarket economic effects on the general population. Neighborhood externalities certainly exist in housing, but internalizing these externalities does not lead to the types of housing programs that have been legislated. The externalities have nothing to do with minimum housing standards for each family. Similarly I find the arguments that education generates externalities unconvincing once one gets beyond elementary education (literacy, etc.).[6] Fire protection is like medical care. Some limited amount of fire protection and code enforcement is necessary to prevent conflagrations, but beyond this a donor has no more interest in his neighbor's fire protection than in his neighbor's fire insurance.[7] Police protection can be described as externality prevention, but it makes more sense to think of the police function as enforcing the rules of the game (a subject to which we shall return).

As a result, while there is a Pareto optimality argument for transfers restricted to particular goods whenever particular goods appear in the donor's utility function (regardless of the cause), this argument does not seem to be a satisfactory justification for actual behavior. What is?

II. The Social Welfare Function

The individualistic social welfare function leads to cash transfers. Interpersonal comparisons that raise or lower individual budget constraints may be made, but as long as individual utilities appear as arguments in the social welfare function, cash transfers dominate restricted transfers. The individualistic social welfare function, however, does not adequately represent social welfare judgments. It ignores the fact that individuals have different levels of preferences and that there is a difference between allowing individual preferences to determine the form of the social welfare function and making social welfare a function of individual utilities.

Individual preferences can be separated into those that concern the ideal rules of the economic game (communism, capitalism, etc.) and the best distribution of economic prizes (the desired degree of equality or inequality) and into those that concern the maximization of personal utility in the current economic game. There is nothing self-contradictory, for example, in seeking to become extremely wealthy and powerful in our current economic game yet believing that a better economic game would be one where there were no "extremely wealthy" prizes to be had.

---

[5] For an explicit example, see the exchange between Gordon Tullock and Edgar O. Olson on subsidized housing.

[6] For an attempt to specify the externalities that flow from education see B. Weisbrod.

[7] Economies of scale arguments do not make fire protection into a public good any more than they make electricity generation into a public good.

To distinguish these two levels of preferences I call one *individual-societal preferences* and the other *private-personal preferences*. This distinction makes it possible to avoid some of the problems inherent in individualistic social welfare functions. In my formulation societies can, if they wish, discuss what constitutes economic equity without having to worry about differences in what, in a loose sense, one might call the efficiency with which people process economic goods (the miser does not need to be given transfers simply because he is a miser and gets more utility per dollar). A preference such as envy is ruled out, not because it does not exist or because it does not affect private personal utility—it does—but because society chooses not to take envy into account in its social rules, even though each one of its members may be envious. In their individual-societal preferences individuals decide to rule out the private-personal preferences of envy, since collectively it can lead to absurd results.[8]

But what are our individual-societal preferences about the structure of the economic game and the distribution of economic prizes? Obviously I do not have the answer to this question, but I can at least say that it is not to be found by analyzing the factors which create private-personal utility. Moreover, when we talk about "the rights of man" or "equality of opportunity," we express individual-societal preferences about the rules and structure of society and the economic game, not private-personal preferences.

Indeed, I would argue that a society's interest in the distribution of medical care springs not from unspecified externalities that affect private personal utility, but from our individual-societal preferences that "the rights of man" include an *equal* "right to life" itself. If you like, we may want to live in a society that is communistic with respect to the "right to life." Because of our private-personal utility blinders, economists act as if it is irrational for a society to desire an equal distribution of the "right to life"; yet it is perfectly rational for society to be communistic with respect to political voting power (one man one vote) or all economic goods and services. The same reasoning that leads us to equality in voting rights leads us to equality in medical care rights. Because of this desire a society may wish not just to establish some minimum standard of medical care, but to equalize medical care—just as there are desires to equalize voting power and not just to insure that everyone has some minimum quantity of voting power. To favor equality in the distribution of the "right to life" and some degree of inequality in the distribution of other economic goods and services is not internally inconsistent. An equal distribution of general purchasing power need not be part of one's conception of the rights of man. *To have equality in the distribution of medical care and still permit inequality in the distribution of other goods and services, it is necessary to provide a voucher or in-kind distribution of medical services.*

Similar arguments could be made for education and housing. Restricted transfers of these commodities spring not from externalities, but from our individual-societal preferences as to what constitutes a good society. Compulsory education implies that the rights of man include the right and *the duty* to acquire some minimum level of education—not because it maximizes the recipient's private-personal utility, but because it maximizes the individual-societal welfare of both the donor and the donee. As a result, restricted transfers flow not from interdependent private-personal utility functions, but from our individual-societal preferences and the resulting social welfare function. In this social welfare function, equality in the distribution of certain "rights"—and, consequently, equality in the distribution of certain specific goals—may be a direct argument.

---

[8] For a longer discussion of the two types of preferences see Thurow (1973).

To determine which goods need to be provided in the form of restricted transfers rather than cash transfers, it is necessary to determine the specific goods that are to be distributed in a manner differing from that of general goods and services. The answer is to be found in the revelation of individual-societal preferences in the political process. Restricted transfers cannot be ruled out, as they would be by an individualistic social welfare function.

III. Limited Consumer Sovereignty

At the heart of the economist's love affair with cash transfers is the doctrine of absolute consumer sovereignty. Everyone is his own best judge of what should be done to maximize his own utility. Real public policies must face up to a modification of this simplistic view—some individuals are not competent to make their own decisions. Historically and legally we have tried to deal with the problem of incompetence by dividing the population into those legally incompetent to make any and all decisions (minors and institutionalized persons) and the rest, who are presumed competent to make any and all decisions.

Increasingly we are coming to recognize that the world is not neatly divided into the competent and the incompetent. There is a continuum of individuals ranging from those who are competent to make any and all decisions to those who are incompetent to make any and all decisions. Half-way houses, drug rehabilitation centers, and their equivalent are being established for criminals, mental patients, the mentally retarded, drug addicts, and the elderly. Obviously it is a difficult problem to establish any individual's degree of incompetence, but the existence of incompetence is a problem that neither governments nor economists can ignore.[9]

Governments have a whole range of public policies that can be used to supplement or supplant consumer sovereignty, but one of the mildest and least coercive of these is the public provision of goods and services in kind. Such in-kind aid can be used to influence individuals to make those decisions that society thinks they would be making if they were "competent."

The extent of the problems created by limited consumer sovereignty is expanded by differences in personal management efficiency. How good is the family in managing and providing the normal housekeeping functions that each family provides in kind to itself? Economists act as if every family's efficiency is both equal and superb. Realistically there is a continuum of family management ability. As a result the same cash income will provide different real standards of living for different families.

Suppose that society has a general income distribution objective that calls for some minimum real standard of living. Does it set its cash transfers at a level where the best family manager can reach the desired minimum, where the average family manager can reach the desired minimum, or where the worst family manager can reach the desired minimum? Given the probable range in management inefficiencies, the cash transfers that would be necessary to give every family at least some minimum real standard of living

---

[9] Governments must be involved since no individual can have the right to declare some other individual's incompetence.

(despite its own inefficiencies) might be impossibly large. As a result, simple efficiency calls for constraining family management inefficiencies with in-kind aid. (Economists deal with efficiency differences between firms by relying on the sacred right of the inefficient to fail, but this is clearly an inadequate response to inefficient families.) One of the key reasons for our general interest in minimum standards of living and cash transfers is our desire to prevent "failing families."

Given a continuum of individuals with varying degrees of competence, transfer systems need a corresponding continuum of transfers ranging from cash, cash with advice, vouchers, in-kind provision, and finally, compulsion. Efficiency may also call for in-kind provision in cases where cash would work but only with large expenditures relative to in-kind aid.[10]

IV. The Creation of Individual Values

The simple model of *homo economicus* fails to describe the process whereby individual wants, preferences, and values are created. Indeed, it asserts that each individual has a set of fixed, innate, fully grown preferences and that values flow from the individual to the society and not in the reverse direction. But modern sociology and psychology suggest that the postulate of static innate wants is simply untenable.

Without thinking about it, every society has always affected the wants of its population, but we must now explicitly recognize that we have always done so,[11] and ask: What individual values should society attempt to create? Americans have always used private and governmental persuasion to create the desired wants (belief in democracy, private property, etc.), but overt discussion of this process is usually associated with the advocacy of totalitarian dictatorships. This association is unfortunate since it obscures the real problem.

The real problem is not whether wants are going to be socially generated (they are), but what systems of overt and covert persuasion are to be allowed. To what extent should the public education system be used to inculcate values? To be against dictatorships does not answer the question. The problem can be put in perspective by imagining the other extreme. What would society look like if there were no social persuasion? With no commonly shared beliefs and wants, chaos would result. Society could not exist. The relevant question is what degree and kind of governmental persuasion should be used to inculcate values. If societies did not believe in creating and enforcing values, laws and police departments would be unnecessary. Everyone would operate on his own values. Enforcing the rules of the game is not simply a matter of force, however. It is also a matter of inculcating the belief that the rules of the game are the right rules.

What role should the public provision of goods and services play in the creation of individual values? Viewed as a problem in literacy, educational vouchers may dominate public schools, but viewed as value creation, public schools may well dominate educational vouchers. Public schools allow society to insure that each individual will be inculcated with, or at least exposed to, the basic values of society. Whether this is necessary or unnecessary obviously depends upon expectations as to what values would be inculcated in private schools. No democracy, for example, could survive a fascist school system.

---

[10] Decision aiders are available in the private market (investment consultants, etc.), but decision-making competence is necessary to purchase a good decision aider, and this is exactly the quality that has been called into question.

[11] The problem is not solved by being against advertising. There are no innate wants for advertising to distort. Wants created by advertising are just as much wants as wants created by any of the other social processes.

Another example exists in the area of drug rehabilitation. A simple application of economic principles would allow a free market in hard drugs, give each person the right to become a drug addict, permit each individual to kill himself if he wished, and interfere only to prevent drug addicts from intruding on the activities of the rest of the population. Yet society seeks to alter the actions of the actual or potential drug addict, not because these are antisocial activities that harm others, but because any society requires positive commitments to its values as well as the absence of negative externalities.

As a result, it is necessary to ask what values the American society holds so deeply and are so fundamental to it that it wants government programs to inculcate these values in each of its citizens. What beliefs—e.g., democracy, free speech, the work ethic—are so basic that society does not want to leave it to the accidental forces of the market place to determine whether individuals are or are not exposed to them? If there are any such beliefs, and I suspect that there are, then it is necessary to ask how the in-kind provision of services can contribute to propagating the desired values. The current disputes over child care are a good illustration of the problem. If public child care is designed to allow each mother an equal opportunity to go to work (regardless of her husband's income), then vouchered private child care centers may be the appropriate answer. If child care is also desired as a method of altering values and characteristics (enrichment is our current euphemism), then public in-kind provision of child care is appropriate. Since society wants to change the values that the mothers would inculcate if they were at home, it obviously cannot allow the mothers to pick the private child care centers that the public funds are to be used to support. In either case, however, restricted transfers dominate cash.

V. Conclusions

Where does this leave us with respect to drawing the line between cash and restricted transfers? While it is not axiomatically true that cash transfers always dominate restricted transfers, the general economic case for cash transfers is strong enough that the burden of proof should always lie on those who advocate restricted transfers. It is their duty to show that the aid is motivated by specific good externalities, that it is part of our individual-societal social welfare function, that it is directed at groups with limited consumer sovereignty, or that it is necessary to the furtherance of some essential value.

With this in mind, I find, for example, the case for in-kind housing assistance unconvincing but the arguments for an *equal* distribution of medical services overwhelming. Indeed, given this framework I suspect that we will find many goods that are now given in-kind that should not be given in-kind, and some goods that are not given in-kind that should be given in-kind.

REFERENCES

H. M. Hochman and J. D. Rodgers, "Pareto Optimal Redistribution," *Amer. Econ. Rev.*, Sept. 1969, *59*, 542–57.

E. O. Olsen, "Subsidized Housing in a Competitive Market: Reply," *Amer. Econ. Rev.*, March 1971, *61*, 220–24.

M. Pauly, "Efficiency in the Provision of Consumption Subsidies," *Kyklos*, 1970, 33.

L. C. Thurow, "The Theory of Grants-In-Aid," *Nation. Tax J.*, Dec. 1966.

———, "Toward a Definition of Economic Justice," *The Public Interest*, Spring 1973.

G. Tullock, "Subsidized Housing in a Competitive Market: Comment," *Amer. Econ. Rev.*, March 1971, *61*, 218–19.

B. Weisbrod, *External Benefits of Public Education*, Princeton 1964.

Earnings Capacity and the Target Efficiency of Alternative Transfer Programs

By IRWIN GARFINKEL AND ROBERT HAVEMAN*

The standard indicator of family economic status is annual money income (*AMI*). The official Social Security Administration (*SSA*) definition of poverty is based on *AMI*, as are analyses of economic inequality among families. The limitations of *AMI* as an indicator of economic position are well known. It fails to incorporate the value of human and physical capital, in-kind public transfers and other public services, intrafamily flows of income and services, and leisure time.[1]

The ranking of families by *AMI* thus may yield a distorted picture of the relative economic status of any given family. One problem of *AMI* is its dependence on the taste for income: a family with a strong taste for income will rank higher than an identical family with a weaker taste. Thus, if the wife in but one of two identical families chooses to work, or if the head chooses to be a craftsman rather than a farmer, that family will rank higher than the other family. In these cases, the difference in ranking reflects a difference in tastes for using economic resources rather than a difference in the level of economic resources.

Evaluation of the equity effects of alternative income transfer programs depends on the measure of the economic status of beneficiaries which is employed. One evaluative criterion—that of "target efficiency"—concerns the extent to which programs focus their benefits on specified target groups.[2] When the target group is the poverty population, and when families are classified as poor or nonpoor by the *AMI* measure, the results of the evaluation may be misleading. Evaluation of program alternatives by this measure may bias choice away from programs which assist families with few economic resources but relatively high tastes for money income.

In this paper, we develop an alternative indicator of family economic position, based not on family *AMI* but rather on a family's ability to generate income when it uses its human and physical capital at capacity. We call this measure *earnings capacity*. This indicator both purges tastes for work from the measurement of economic position and abstracts from temporary fluctuations in income.[3] The compo-

---

* Associate professor of social work and staff member of the Institute for Research on Poverty, University of Wisconsin–Madison, and professor of economics and director of the Institute for Research on Poverty, University of Wisconsin–Madison, respectively. David Betson assisted with the paper.

[1] Efforts to develop a more comprehensive indicator of a family's command over goods and services include: James Morgan, Martin David, Wilbur Cohen, and Harvey Brazer; Burton Weisbrod and W. Lee Hansen; and Michael K. Taussig.

[2] The concept "target efficiency"—the percent of program benefits to pretransfer poor families—was suggested by Weisbrod.

[3] Such fluctuations can substantially influence a family's ranking in the money income distribution. By

sition of families at the bottom of the distribution of family earnings capacity—the earnings capacity poor—is identified and compared to the poverty population as defined by the standard definition. Finally, we compare the target efficiency of several alternative income transfer programs using both the officially defined poverty population and an equivalent number of families at the bottom of the earnings capacity distribution as the target populations.

The principal objective is to ascertain the extent to which estimates of target efficiency based on *AMI* may be misleading if the objective is to assist families with low levels of marketable skills and few physical assets. The emphasis on target efficiency in this study does not imply that this equity criterion is or should be the primary criterion for choice among transfer policy alternatives. Such evaluation rests on a variety of criteria—taxpayer cost, benefit "adequacy," incentives for alteration in work effort and family size and structure, as well as target efficiency. However, if the targeting of benefits on families with few human and nonhuman assets is an objective, the measurement of target efficiency should not be based on an indicator of poverty with the conceptual flaws of *AMI*.

I.

Earnings capacity reflects the ability of a family—given its current endowment of physical and human capital—to generate a net income flow if the family uses that endowment at capacity. In this study, an estimate of family earnings capacity was developed for each of 50,000 families in a weighted national sample of families.[4] In measuring the earnings capacity of human capital, we estimated separate earnings functions for men and for women and used them to impute the full-time, full-year earnings of an individual on the basis of his or her demographic characteristics. The independent variables included were the standard measures of human capital: *age*, taken to reflect job experience and training as well as the obsolescence of skills and physical and mental capacities with more advanced years; *years of schooling*, to reflect the stock of human capital conveyed by education; *race*, to reflect labor market discrimination among blacks and whites; *marital status*, to reflect the higher incidence of physical and mental disabilities and other characteristics leading to lower labor productivity among single than married persons;[5] *location*, to capture real productivity differences not captured by the human capital variables;[6] and *weeks worked* and *part-time-full-time* employment status. On the basis of results from previous studies, interactions were introduced to capture the differential effects of age and education on earnings by race and the differential effects of education on earnings

---

eliminating the effects of income instability, the earnings capacity concept is akin to, though not identical with, permanent income. In its pure form, permanent income reflects the present value of expected lifetime consumption. Earnings capacity takes as given the stock of human and physical capital at a point in time and estimates the return accruing to capacity use of these assets. Conceptually earnings capacity is more closely related to Gary S. Becker's notion of "full income."

[4] The sample was that of the 1971 Current Population Survey, "aged" to be representative of the 1973 *U.S.* population. The extract tape was developed by the Urban Institute. The observations on the survey were also modified so as to yield a national unemployment rate of 4.9 percent. For a more detailed discussion of the adjustment of the 1971 data, see Nelson McClung, John Moeller, and Eduardo Siguel.

[5] For women, a dummy variable to reflect *child status* was also used, reflecting the association of work experience with both marital status and child status—single women without children are more likely to have work experience than married women, and married women without children are more likely to have recent work experience than those with children.

[6] The location variable, unlike the other human capital variables, may be considered as much a choice as a predetermined variable. While location may be altered over the long run, at any point in time the labor market in which a family is located is a determinant of its earnings capacity.

by age. The results of these regressions were as predicted by the human capital model.[7]

On the basis of the estimated coefficients, an individual's human capital characteristics, and the assumption that each individual worked at capacity, each family head and spouse in the sample was assigned an earnings capacity—EC_h or EC_w.[8] However, some individuals fail to exploit their earnings capacity. Hence, this imputed value was adjusted to the extent that time away from work was attributable to involuntary limitations.[9] To the adjusted earnings capacity from human capital is added the capacity return on physical assets. While this capacity return is defined as the income flow attributable to full use of these assets, data limitations required that the actual flow of income from interest, dividends, and rent be used in the estimation of this component of family earnings capacity.[10] Finally, to arrive at a concept of *net* earnings capacity, necessary expenses attributable to the use of human capital must be subtracted from the full earnings capacity of a family. In the empirical illustration of the earnings capacity concept presented here, minimally acceptable child-care costs are the only required expense deducted.[11] The empirical measure of net earnings capacity (NEC) based on these assumptions is thus defined as:

$$NEC = EC_h\left(\frac{50 - W_{su}^h}{50}\right) + EC_w\left(\frac{50 - W_{su}^w}{50}\right) + Y_1 - CC$$

where

EC_h = head's imputed annual earnings capacity at 50–52 weeks of full-time work

EC_w = wife's imputed annual earnings capacity at 50–52 weeks of full-time work

W_{su} = weeks sick or unemployed

Y_1 = income from interest, dividends, rents, alimony, and miscellaneous other sources (not including government transfers)

CC = minimally acceptable annual child-care costs taken to be $1,510 per child five years of age or less and $376 per child six to fourteen years of age.

For several reasons, then, the estimate of NEC based on these assumptions may deviate from the pure earnings capacity concept. The estimating procedure fails to capture all of the determinants of earned income (for example, motivation, *IQ*, detailed physical characteristics, and chance); the estimation of the capacity return from physical assets is incomplete; only a portion (albeit the largest portion) of required expenses is deducted from capacity work effort. While these deficiencies may lead to a sizable misestimation of pure earnings capacity for any individual or family, the measure used here should not be far from the mark for demographic groups. Moreover, and in contrast to the *AMI* measure of economic status, it is independent of relative preferences for money income.

---

[7] The specification of the regression models and the regression coefficients and statistics are available from the authors upon request.

[8] For each individual "50 weeks per year" was assigned as the weeks worked variable and "full-time work" was assigned as the full-time-part-time variable.

[9] These adjustments for illness, disability, and unemployment build some temporary reductions in income into the earnings capacity measure. For example, the data did not indicate whether a reduction in employment was due to either a permanent disability and/or a temporary illness. While unemployment or illness may represent a temporary reduction in an individual's earnings capacity, from an aggregative point of view both temporary illnesses and unemployment are a permanent phenomenon.

[10] This procedure implicitly assumes that nonhuman capital is being used at capacity. To the extent this is not true, the earnings capacity of some families is understated. Further understatement would result to the extent that the return from some assets is not recorded as money income and cannot be imputed from the available data. This is the case with the return from owner occupied housing in the survey data used in this study.

[11] This sole deduction of required expenses from full earnings capacity yields an overestimate of the discretionary income available to a family from capacity use of its resources. Additional analyses of the sort discussed below are also being done using a gross earnings capacity concept which neglects the deduction of required expenses. Other analyses use alternative specifications of the child-care cost function, including a function which recognizes the existence of economies of scale within the family. The estimated costs for the minimally acceptable child care are inflated 1968 estimates from B. Bernstein and P. Giacchino.

II.

By placing both the earnings capacity and *AMI* measures over the 1973 poverty line for the family—that is, by obtaining both an *AMI* and earnings capacity "welfare ratio" for each family—one can compare the composition of the *AMI* and earnings capacity poverty populations. About 15 percent of the nation's families in 1973 were poor in *AMI* terms. If this figure is accepted as a benchmark, an equivalent number of families with the lowest *NEC* welfare ratios can be isolated and the composition of the two groups compared. If earnings capacity is regarded as a superior indicator of economic status than *AMI*, the composition of those families with the lowest earnings capacities can be taken as a more accurate description of the characteristics of the poor than that based on the official statistics.

Perhaps the most significant difference between the composition of the two groups concerns the labor force participation of family heads.[12] While about 35–40 percent of the lowest earnings capacity family heads work forty-eight or more weeks per year, only about 20 percent of the *AMI* poor are full-year workers. This is reflected in the fact that nearly two-thirds of the families designated as poor by the standard definition worked less than thirteen weeks per year.

While the proportion of poverty families which are female-headed families with children is approximately the same for the standard and earnings capacity definitions —52 to 54 percent—the factors which account for this high percentage are quite different in the two cases. In the standard poverty definition, a large number of female-headed families with children are counted as poor because the family head does not work.[13] In contrast, such families form a high proportion of the earnings capacity poor not because they do not work, but because, relative to the rest of the population, they can earn so little even if they work at capacity.[14]

Several other interesting comparisons emerge. First, the incidence of earnings capacity poverty among blacks is somewhat higher than *AMI* poverty—the percent of blacks in the bottom 15 percent of families in the two distributions is 28 percent and 22 percent, respectively. Second, the earnings capacity measure indicates a much higher incidence of poverty among families whose head is below fifty years of age than does the standard measure—62 percent to 43 percent. In terms of the education and occupation of the family head, the earnings capacity measure indicates a higher proportion of poverty families with less than a high school education —95 to 87 percent—and employed in lower status, lower skill occupations than does the standard definition. About 50 percent of poor households by the standard definition are unrelated individuals, compared to 17 percent of earnings capacity poor households. Similarly, while over 40 percent of the earnings capacity poor have five or more family members, only about 17 percent of the *AMI* poor are families of this size. Finally, the percent of *AMI* poor with farmer, farm laborer, and household worker occupations is substantially higher than for the earnings capacity poor.[15]

---

[12] Detailed tabular comparisons of the composition of the two groups are available from the authors upon request.

[13] They are poor by the standard definition in spite of the fact that AFDC payments are targeted on this group of families.

[14] This finding becomes clear through the use of a regression analysis with a 0–1 dependent variable representing poverty-nonpoverty status of individual families and the independent variables consisting of family characteristics. When weeks worked is held constant, female headship contributes much less to the probability of being poor by the standard definition than by the earnings capacity definition.

III.

Recent years have witnessed substantial national debate regarding the pros and cons of alternative income maintenance proposals. The debate on proposed changes in the transfer system usually focuses on the efficiency and equity impacts of these plans. One persistent equity question raised about income maintenance alternatives concerns their "target efficiency"—how are program benefits distributed among the beneficiaries, i.e., to what extent are they targeted on those who have the greatest need of them? Heretofore, estimates of target efficiency have employed a concept of economic status based only on *AMI*. But it is now possible to evaluate the effectiveness with which a range of transfer program alternatives target their benefits on families with the lowest economic position by both the conventional and the earnings capacity definition proposed here.

Table 1 provides simulation estimates of the target efficiency of seven income transfer programs on both the earnings capacity and *AMI* basis. Five of the programs (the negative income tax, the earnings supplement, and the two wage rate subsidy programs) are general proposals which are not embodied in specific legislation or administration sponsored bills. The H.R.1 plan for families was an administration proposal, debated at length but not passed by the Congress; the Aid to Families with Dependent Children program (*AFDC*) is an existing program. With the possible exception of the negative income tax proposal, none of these plans would be intended as income support programs for the aged; consequently, the target efficiency and cost estimates are for the nonaged population.

The *AFDC* program provides assistance primarily to female-headed families with children.[16] Eligibility standards and benefit schedules are set by state governments.[17] The simulated distribution of the benefits and the simulated total transfer cost—$3.6 billion—of this program is based upon the responses of recipients in the Current Population Survey.[18]

H.R.1 would have provided federal benefits of $800 per adult and $400 per child (for the first six children). The first $720 of earned income would have been subject to a zero tax rate; earnings in excess of $720 per year would have been taxed at a rate of 67 percent. Unearned income would have been taxed at a 100 percent rate. While H.R.1 would have replaced the current *AFDC* program, the bill contained a provision for state supplementation of basic federal benefits to insure that no current welfare beneficiary would have experienced reduced money income because of the program. The net increment of budgetary costs attributable

---

[15] A larger incidence of income underreporting in these occupations could contribute to this result.

[16] Approximately 85 percent of the recipients of Aid to Dependent Families benefits are female-headed families.

[17] See Irene Lurie for a description of the benefit structure of the AFDC program.

[18] The Current Population Survey captures about 75 percent of public assistance payments. See Dorothy Projector and Judith Bretz. Our target efficiency estimates assume that unreported transfers are distributed among families in the same way as reported benefits.

TABLE 1—TARGET EFFICIENCY MEASURES OF ALTERNATIVE INCOME TRANSFER POLICIES, BY CURRENT INCOME AND EARNINGS CAPACITY

	Percent of Total Benefits Accruing to Pre-Program Poor		
	Current Income Poor	Earnings Capacity Poor	Total Cost $ Billion
AFDC	88	92	3.6
H.R.1—Family Assistance Plan	84	54	4.6
Negative Income Tax	100	52	4.1
Earnings Supplement—Family Earnings (ES_1)	72	52	3.7
Earnings Supplement—Head's Earnings (ES_2)	61	56	3.9
Wage Rate Subsidy—Head and Spouse (WR_1)	39	29	6.2
Wage Rate Subsidy—Head Only (WR_2)	64	44	3.3

to this package of policy changes is $4.6 billion.[19]

The negative income tax plan which is simulated has a guarantee of $800 per adult and $400 per child (for the first six children) and a tax rate of 55 percent on both earned and unearned income.

In an earnings supplement plan, increments to earned income are subsidized up to some stated earnings level; beyond that level, increments to earned income result in decreases in the total supplement. For earnings below the stated level, a negative tax rate (or a subsidy rate) exists; earnings above that level are subject to a benefit reduction rate similar to that of a negative income tax.[20] Estimates of the target efficiency of two variants of an earnings supplement are presented below. In the first earnings supplement plan (ES_1), supplemental payments are based on *family* earned income. The supplement rate is 50 percent up to earnings of about $1,875 (for a family of four). Increments to earnings beyond that level reduce total supplements by 30 cents per dollar earned. Unearned income is subject to a zero tax rate. For a family of four the break-even level is $5,000, varying by $900 per family member for families larger or smaller than four.[21] The total cost of the program is $3.7 billion. The second earnings supplement program (ES_2) is identical to the first except that payments are based only on the head's earnings. The cost is $3.9 billion.

In a wage rate subsidy plan, income support is based on an individual's wage rate and the number of hours worked. Eligible individuals with wage rates below some break-even wage rate are entitled to a subsidy for each hour worked. Given the number of hours worked, the total benefit depends on the subsidy rate, which is the percentage amount of the difference between the individual's own wage rate and the break-even wage rate of the plan. The target efficiency of two variants of a wage rate subsidy is presented in this paper. In both plans the break-even wage rate is conditioned on family size. For a family of four, the break-even wage is \$2.30 per hour; incremental family members (up to eight) alter the break-even wage by 25 cents per hour per family member. The subsidy rate is 50 percent. In the first wage rate subsidy plan (WR_1), all family heads and spouses with wage rates below the break-even wage rate are eligible for a subsidy.[22] In the second plan (WR_2), the subsidy is based on the wage rate of the family head alone. Because self-employed individuals are not likely to be eligible for wage rate subsidy programs due to severe administrative and policing problems, the target efficiency and cost estimates of WR_1 and WR_2 are for the non-self-employed population. The cost of these plans is \$6.2 and \$3.3 billion, respectively.[23]

---

[19] The simulation program for the detailed H.R.1 program was designed by the Urban Institute, as part of their transfer program simulation effort.

[20] For a discussion of the characteristics and impacts of such a program, see Robert Haveman. A variant of this strategy is implicit in the recent proposal of the Senate Finance Committee.

[21] The break-even level ceases to increase with family size at eight members.

As Table 1 demonstrates, the *AFDC* program possesses high target efficiency, irrespective of the measure of economic status employed. Indeed, the target efficiency of this program is slightly higher under the earnings capacity than under the *AMI* measure. While 88 percent of the total *AFDC* benefits go to family units in the bottom 15 percent of the *AMI* distribution—the official poor—92 percent of total benefits go to family units in the bottom 15 percent of the earnings capacity distribution—the earnings capacity poor. Given our earlier finding that the percentage of poor families with children headed by females was nearly the same for both definitions, the similarity of these estimates is not surprising.

While H.R.1 is nearly as target effective as the *AFDC* program by the *AMI* measure, it is much less effective in targeting its benefits on the earnings capacity poor. Only 54 percent of the total benefits go to the bottom 15 percent in the earnings capacity distribution. This result indicates that a significant number of families in the bottom 15 percent of the *AMI* distribution who would be aided by H.R.1 are either only temporarily poor or, for some other reason, have earnings capacities in excess of their actual earnings.[24] These reasons could include relatively high tastes for either leisure or for particular kinds of employment with low remuneration or the existence of some characteristic not captured in the measurement of earnings capacity which constrains successful market activity.

The most interesting comparisons are those among the modest negative income tax plan, the two earnings supplement plans, and the two wage rate subsidy plans. Almost 100 percent of the benefits of the negative income tax plan go to the *AMI* poor. The current income target efficiency measures of the earnings supplements and wage rate subsidies are substantially lower—72 percent for ES_1, 61 percent for ES_2, 39 percent for WR_1 and 64 percent for WR_2. However, using earnings capacity as the indicator of economic status, the earnings supplement programs appear slightly more target efficient than the *NIT* program, although the difference is small. Thus, while there may be efficiency and equity grounds for preferring a negative income tax program to an earnings supplement, these estimates indicate that—if earnings capacity is preferred to current income as an indicator of economic status —the difference in target efficiency is not one of them.

---

[22] The simulation of the wage rate subsidy plan was done on the data from *A Panel Study of Income Dynamics*, conducted by the Institute for Social Research, University of Michigan. Hourly wage rates are available in this data source but not in the Current Population Survey (CPS). Simulations of the NIT program discussed above on the Panel Study data yielded similar results to those on the CPS.

[23] While the costs of these proposed strategies (*NIT*, *ES*, *WR*, and H.R.1) are not identical, they are of comparable magnitude. The relationship of the *AMI* to earnings capacity target efficiencies for any one of these programs would not be altered in any significant way if adjustments are made to insure equal cost of programs.

[24] To the extent that the earnings capacity measure is adjusted for illness and involuntary unemployment, these factors do not contribute to this result.

Similarly, the earnings capacity target efficiencies of the wage rate subsidy confined to family heads (WR_1) and the modest negative income tax are relatively close—54 percent versus 44 percent. Only the wage rate subsidy based on both head and spouse earnings (WR_2) is substantially less target efficient in terms of earnings capacity than the other programs. Benefits from this plan accrue to many families in which the spouse earns a low wage, while the family head commands a relatively high wage rate.

A final point of interest concerns the reversal in target efficiency of the two earnings subsidy plans in moving from the *AMI* to the earnings capacity measures of economic status. ES_1, in which benefits are based on the sum of husband's and wife's earnings, appears to be the more target efficient of the two under the *AMI* measure. Conversely, ES_2, in which benefits are based on the husband's earnings only, is more target efficient under the earnings capacity measure. Some families will have incomes in excess of the *AMI* poverty level by virtue of the spouse's earnings. A portion of these families will also have incomes in excess of the break-even level of earnings in ES_1 by virtue of the wife's earnings. Because ES_2 ignores the wife's earnings in calculating payments and eligibility, families who are not in the bottom 15 percent of the *AMI* distribution will receive benefits from ES_2 but will not receive benefits from ES_1. As a consequence, ES_2 will be considered less target effective than ES_1 by the *AMI* definition of poverty. However, inclusion of a family in earnings capacity poverty is not dependent on the work status of the spouse, hence, the small difference in the target efficiency of the two plans when earnings capacity is employed as the indicator of economic status.[25]

IV.

In this paper, we developed an earnings capacity measure of economic status. A comparison of the lowest 15 percent of families in the welfare ratio distributions based on *AMI* and earnings capacity—that is, the official poor and the earnings capacity poor—yielded interesting findings. In both distributions, for example, female-headed families constituted an equally large percentage of the bottom 15 percent of families, indicating that even if a substantial number of female heads worked much more than they now do, they would still be poor. In addition, we found that working poor families constituted a significantly larger percentage of the total earnings capacity poor than of the *AMI* poor, reemphasizing the inequity in the failure to provide assistance to the working poor.

A comparison of the *AMI* to the earnings capacity target efficiencies of alternative income transfer programs was also illuminating. The *AFDC* program was seen to be highly target effective for both measures. However, the greater *AMI* target efficiency of negative income tax programs relative to earnings subsidy programs was virtually eliminated or even slightly reversed when target efficiency was measured in terms of aid to the earnings capacity poor. Similarly, the gap between the target efficiency of the negative income tax and wage rate subsidy plan for family heads was substantially reduced when target efficiency was measured in terms of aid to the earnings capacity poor. Only the wage rate subsidy plan for heads and spouses remained equally target ineffective relative to other programs when earnings capacity rather than *AMI* was used to define poverty.

---

[25] Whether the earnings supplement based on head's earnings will be more or less target effective by earnings capacity than a plan with otherwise identical parameters except that payments depend on the earnings of head and spouse depends on the parameters of the plan. However, in every case of several simulated, the former appeared more target efficient than the latter when poverty was defined in terms of earnings capacity.

Our results should not be interpreted as an argument that, in general, earnings subsidies are superior to negative income tax programs, or that wage rate subsidies for family heads only are preferable to wage rate subsidies for all family members. As noted above, target efficiency is only one of many criteria by which income transfer programs should be judged. For example, whether to confine eligibility for wage rate subsidies to family heads depends not only on the target efficiency of this policy, but also on the incentives created by such a policy for alterations in family structure. Similarly, the choice between negative income tax, earnings supplements, and wage rate subsidies depends upon the transfer and administration costs, benefit adequacy, and work and family incentives, as well as the target efficiency of the alternative programs. These estimates are not a full evaluation of the desirability of alternative transfer programs. They do, however, illuminate the shortcomings of both the official definition of poverty and of the target efficiency evaluation criterion when based on the official definition.

REFERENCES

G. S. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493–517.

B. Bernstein and P. Giacchino, "Costs of Day Care: Implications for Public Policy," *City Almanac*, August 1971.

R. Haveman, "Work Conditioned Subsidies as an Income Maintenance Strategy," in Joint Econ. Comm., *Studies in Public Welfare*, paper no. 9, 1973.

M. Krashinsky, "Day Care and Welfare," *Stud. in Public Welfare*, Joint Econ. Comm., Washington 1973.

I. Lurie, "Tax Rates in the *AFDC* Program" *Nat. Tax J.*, forthcoming.

N. McClung, J. Moeller, and E. Siguel, *Transfer Income Program Evaluation*, working paper 950–3, Urban Institute, Washington 1971.

J. Morgan, M. David, W. Cohen, and H. Brazer, *Income and Welfare in the United States*, New York 1962.

D. Projector and J. Bretz, "Measurement of Transfer Income in the Current Population Survey," paper prepared for the Conference on Research in Income and Wealth, 1972.

M. K. Taussig, *Alternative Measures of the Distribution of Economic Welfare*, Princeton 1973.

B. Weisbrod, "Collective Action and the Distribution of Income: A Conceptual Approach," in U.S. Congress, Joint Economic Committee, *The Analysis and Evaluation of Public Expenditures, The PPB System*, Washington 1969.

——— **and W. Lee Hansen**, "An Income-Net Worth Approach to Measuring Economic Welfare," *Amer. Econ. Rev.*, Dec. 1968, 58.

Conflicting Objectives in Income Maintenance Programs

By ROBERT I. LERMAN AND ALAIR A. TOWNSEND*

For some years many economists have agreed that the way to bring order to the income maintenance chaos is to adopt some form of negative income tax (*NIT*). Just five years ago, the income maintenance system was simpler and less expensive. It was easier to propose "good" simple reform plans. Since then the world has become increasingly complex: the food stamp program has been transformed into a near-universal income guarantee; a federal cash guarantee has been enacted for the aged, blind, and disabled; and expenditures, caseloads, and benefit levels under cash and medical welfare programs have risen substantially.

Because of this welfare explosion, any program rationalization—even if it replaced much of the hodgepodge of current programs with one generous program—would hurt financially many more recipients today than in 1968. And more recipients are enrolled in programs whose benefits decline with income, adding to the problem of high, cumulative, marginal tax rates.

Thus, it is more difficult now to design an *NIT* that is technically and politically appealing. Many of those who supported the *NIT* as a way of both rationalizing the income maintenance system and helping the poor are looking to other kinds of programs for the answer. There is an upsurge of writing to the effect that in-kind transfers may not be so bad, and that, since *NIT* has proven politically unpopular, we should move towards programs which emphasize work—such as wage subsidies and public employment. Unfortunately, none of these alternatives escapes from the technical difficulties of the *NIT* because they all create adverse incentives—with respect to family structure, work behavior, and childbearing—and difficult equity choices.

This paper briefly outlines the growth in programs since 1968, then highlights design problems and necessary tradeoffs present in any income maintenance system, and finally points out some difficult transition problems in trying to achieve the dual objectives of rationalization of the system and providing more income for the poor. Our theme is that there is no easy way out.

I. While We Weren't Looking

Federal and state income maintenance expenditures rose $54 billion ($44 billion in constant 1972 prices) from fiscal year (*FY*) 1968 to *FY* 1973 (see U.S. Budget). But between 1968 and 1972 the poverty gap remained at about $12 billion (at 1972 prices) and the poverty population dropped only slightly to 24.5 million persons.

Why did the extra $54 billion have so little apparent impact? If we had had a mandate to spend an additional $54 billion on income maintenance in 1973, we almost certainly could have eliminated poverty and "reformed" welfare as well. But most of the $54 billion was spent on social insurance benefits (such as social security and unemployment insurance) and benefits in kind. Social insurance expenditures increased from $25 billion to $55 billion. Much of the expenditure on social insurance, of course, goes not to the poor, but to middle-income persons. For example, of all unemployment insurance benefits paid in 1970, 15 percent went to the 28 percent of families with incomes under $5,000 while 28 percent went to the 29 percent of families with incomes above $15,000. (See Martin Feldstein.) Such programs as social security do help lift many people out of poverty, but for many of the poorest they merely replace public assistance dollars that would otherwise be spent. Thus, massive increases in social insurance benefits fail to make a massive impact on poverty.

---

* Staff economist and research director, respectively, Subcommittee on Fiscal Policy, Joint Economic Committee, U.S. Congress. These views are ours and do not necessarily reflect those of the Subcommittee on Fiscal Policy, the Joint Economic Committee, or the staff. We thank Robert Harris, Vee Burke, and James Storey for their helpful comments.

Food, health, and housing benefits in kind also more than doubled, rising to $21.2 billion. Although these helped low-income people, the actual gain in real income is not easily measurable and the impact on measured poverty was zero because these benefits are not counted as income by the Census Bureau.

This pattern of proliferating programs and burgeoning expenditures could and probably will continue for some time without bringing desirable welfare reform or eliminating poverty.

II. What Kind of Income Maintenance Program Should We Seek?

What is the ideal income maintenance system? This question has received much attention recently; most answers to it give primary emphasis to work incentives and have much less to say about other conflicts that a sensible income maintenance system must resolve.

In assessing policy alternatives, one often finds the economic principles of efficiency and of horizontal and vertical equity handy but not sufficient. That tradeoffs must be made is well known. But quantitative estimates of these tradeoffs are either notoriously unreliable or totally unavailable. We know very little about the budget cost and impact on labor supply of different tax rates, or the effect on family stability of different kinds of benefit schedules.

The notion of equal treatment of equals raises additional problems. In the dynamic world in which we live, one must apply the equity test to both opportunities and results. Do observed differences in family size, family structure, income flow, level of work effort, and location essentially represent differences in taste, or do they instead reflect circumstances over which individuals have little or no control, e.g., constraints? Should one-parent families be favored over two-parent families based on their current disadvantages or receive similar treatment based on equal access to the same family structure? Thus, regardless of whether adverse incentives actually affect behavior, difficult normative questions remain regarding the direction of redistribution.

Objectives concerning family structure, family size, and household composition are inevitably in conflict. A program should not institute rewards for family splitting and penalties for marriage or sharing households, and it should not offer substantial incentives for having large numbers of children. Unfortunately, even these few goals cannot be achieved simultaneously.

Next, consider the problem of defining the recipient unit. Should we pay benefits and count income on the basis of the individual, the family, or the household? (See W. A. Klein.) Such choices are important in the construction of public employment and wage subsidy plans as well as *NIT* plans, since eligibility for most employment-based proposals depends on income.

Under a system which treated each individual as a recipient unit, guarantees would be independent of family or household formation. Thus, marriage would not be discouraged and poor persons could economize by sharing households without a financial penalty. But if all individuals were eligible regardless of age, such a program would offer an undesirably large financial incentive for having children. Because it would take no account of economies of scale in family size, it would redistribute too much to large families relative to small ones. Moreover, an individual definition would ignore actual income sharing of all kinds, including sharing within the family.

A system which treated the family as the recipient unit could take account of the economies of family size and of income sharing. However, it would create a financial disincentive for men and women living together to marry. Total benefits probably would be smaller for a family unit which included the man and his income than for a recipient unit composed of the woman or the woman and children only. Any employment or wage-based program which divides the population into "employables" and "unemployables" and retains a welfare-like program for the latter group, especially women with children, will face the identical problem when using the family as recipient unit (see Arnold Packer).

Definition of the recipient unit as the household would eliminate adverse family formation incentives by treating people living together in the same way whether or not they are related. However, this policy could discourage economies achieved when households are shared. A low-income family could lose its benefits if it moved in with a moderate-income family. Such a definition also implicitly presumes the existence of income sharing among household members in the absence of legal responsibility or any evidence of sharing. Finally, household definitions are probably the most difficult to administer because of the problems in identifying unit membership and the more frequent changes in unit composition.

Payment differences by geographic areas involve another tradeoff. Since costs of living vary by location, assuring equal living standards for those with equal incomes would require variations in payment levels. However, geographic payment differentials are inefficient. By allowing the lowest income groups largely to ignore the high real costs in areas with high prices, welfare payment differentials interfere with the allocative efficiency of the price system. As Irene Lurie points out:

> The price variation within regions is larger than the variation among them. Price-level adjustments between closely situated areas would adjust for a great deal of the total variation, but it is precisely these adjustments which could most easily influence migration. In dividing the country into the price-level areas, therefore, there is a tradeoff between the accuracy of the price-level adjustment and the minimization of its effect on migration. [p. 246].

Even the equity basis for geographic differentials is not particularly persuasive. For example, costs are high in San Francisco partly because people enjoy its environmental advantages. Providing higher welfare payments for San Franciscans would mean that low-income residents would gain the locational advantages without suffering an appreciable loss in other goods and services. Further, there is the question of equity between recipients and nonrecipients. Middle- and upper-income people in general receive no similar tax advantage from choosing a high cost and high benefit location.

Another issue that forces choices among competing goals relates to the accounting period, the time period over which income is counted and payment levels are computed. Short accounting periods allow government payments to be most sensitive to current needs at the cost of reduced equity over the long run and reduced savings incentives. Again, at the root of the conflict is the extent to which a family has control over its consumption flow. Those who favor accounting periods of one month or one quarter believe that low-income people have little ability to save, so that above-average income in a previous month or quarter is unrelated to the family's consumption ability in the current month or quarter. Advocates of longer accounting periods maintain that people do have the ability to save and, often, the ability to regulate their periods of work. It is well established that farmers or other self-employed persons can affect their measured within-year net income flow by accounting devices. Thus, families capable of earning the same annual incomes have the ability to achieve the same consumption levels at each point in the year and, therefore, ought to receive the same benefits.

A related issue is that the adoption of short accounting periods increases the pressure for assets tests. Assets tests, which exist in most current programs, restrict eligibility to those with assets below a specified level. As the accounting period becomes shorter, current need becomes more dependent on asset levels in that asset decumulation can more easily satisfy needs. As an extreme example, note that income on a particular day is a poor indicator of current need; one also has to know about accumulated savings. The necessity for considering assets does not require that we continue the current policy of paying benefits to those slightly below the asset limit and denying them to those slightly above. This policy is highly inequitable. Moreover, it discourages savings and encourages asset decumulation. Unfortunately, the more rational policy of imputing income to assets has gained little acceptability in policy circles, and indeed it only makes sense if long accounting periods are used.

III. Obstacles to Reform

Awareness of the many conflicting objectives should remind us that perfection is out of reach, but should not discourage us. But transition to a better system is made particularly difficult by the poor integration and poor design of existing income transfer programs. This section examines obstacles in the path toward generally desirable goals.

A. *Growth of In-Kind Programs*

The rationale for the proliferation of in-kind transfers is that different programs are designed to deal with different "needs," so transfers come in the form of subsidized housing units, groceries, school lunches, day care, medical services, and higher education. These programs are inefficient in two senses. First, they are inefficient in that they impose constraints on consumers' efforts to maximize their utility functions. Second, they are budgetarily inefficient in that costs of government-provided goods or services are often higher than market costs. For example, the U.S. Department of Housing and Urban Development estimates that it must spend almost $2 to provide to public housing tenants services that would be valued at $1 in the market (pp. 4–82).

One might expect that the transition from in-kind to cash benefits would come easily, since recipients would probably prefer to take less in cash than the present amount of government expenditures per recipient, and the government should prefer to spend less and achieve the same level of welfare. However, shifts from one program to another rarely prevent interest groups from trying to hold on to the old program while enacting the new. An example is the recent aborted attempt to convert food stamps to cash for aged, blind, and disabled welfare recipients. Not only do special constituencies promote such in-kind programs as school lunches and day care, but a transition to cash usually threatens an income loss for some recipients. This is because current in-kind programs contain more inequities than we could tolerate in a cash replacement program. Spreading around the same expenditures more evenly would mean welfare losses for many current recipients (see J. Storey, A. Townsend, and I. Cox.).

B. *The Cost of Greater Horizontal Equity*

The present distribution of benefits is haphazard, making horizontal equity nearly impossible to achieve if one of the rules of reform is (as it seems to be) that no one should lose from the change. Currently, some families receive benefits from many programs while neighbors with similar incomes receive little or nothing. These variations in benefits occur primarily because of categorical eligibility rules and closed-end appropriations. The most significant groups excluded for categorical reasons are poor and near-poor families with both parents present. They are far less likely to benefit from Aid to Families with Dependent Children (*AFDC*), medicaid, and day-care services than are low-income, one-parent families. The common pattern of legislating easy eligibility for program benefits and mandating high service standards while appropriating little money necessarily excludes most eligible families from such benefits as subsidized housing and day care.

A survey undertaken by the General Accounting Office (*GAO*) for the Subcommittee on Fiscal Policy compiled program data on households in six low-income areas. According to this survey, 60 percent of the households received at least one benefit and 11 percent received at least five benefits. Of those households receiving at least five benefits—a nontrivial 20 percent of all beneficiaries—average cash plus in-kind benefits in a western city equaled $144 per month per person or $6,722 per year for a family of four. Average private income per four-person unit for this group was $160, or $1,916 per year. Counting only cash, food, and housing benefits for those with no private income and five or more benefits, the *GAO* found average benefit levels on an annual basis varied from $3,336 in a southern Atlantic city to $5,112 in an eastern city (Storey, Townsend, and Cox).

Estimates made by Benjamin Okner show how high positive income tax rates would have to be to support guarantees of this magnitude. Ignoring possible labor supply reductions and without restructuring social security, but with elimination of expenditures on food, housing, and cash public assistance programs, and assuming enactment of a partial tax reform plan, Okner calculated that a demogrant that guaranteed $3,600 to a family of four in 1970 could have been financed by a proportional income tax of 46 percent. Thus, most families would pay higher marginal, but lower average tax rates. Use of an *NIT* plan could reduce positive tax rates but would require higher negative tax rates. To support a demogrant in the $5,000 guarantee range, marginal tax rates would have to be well over 50 percent. And $5,000 is lower than benefits actually received by a large number of households.

The conclusion is straightforward. Benefits for many households are already too high in the sense that such levels cannot be extended to all eligible families without disastrously high positive and/or negative tax rates. So, income maintenance programs—whether an *NIT*, a public employment, or wage subsidy program—will require either that some families suffer reductions in benefits or that families with similar incomes continue to receive markedly different benefits. A common proposal for resolving this conflict is to allow current recipients to retain their higher benefits while phasing in the major reform plan. Unfortunately, the almost daily proposals for new in-kind proposals, such as expanded day care and college grants and loans, hamper this strategy by making it even harder to keep a lid on current benefits long enough for the reform plan to catch up.

Suppose we ignore the problem of reducing benefits for current recipients and assume there is an acceptable way of providing "grandfather" (no-loss) rights. How do we handle the issue of geographic payment differentials and the associated issue of whether to allow state supplementary payments? One solution is a federally financed program—whether *NIT*, wage subsidy, or public employment—with no geographic differentials or, at most, differentials related to cost-of-living differences, but not to standard-of-living differences. This policy would tend to transfer money from moderate- and high-income areas to poor areas. While this result is economically desirable, it raises political problems. Already, leaders in New York, Connecticut, and New Jersey have complained that the national standards in the food stamp program mean lower average benefits in their states than in other states.

A second possible approach would be a program with a federally financed benefit floor in which states could supplement benefit levels with their own funds. (This policy has recently been put into effect in the new Supplemental Security Income program for the aged, blind, and disabled.) Under this approach wide payment variations would persist, with little relation to cost-of-living differences, although the current variation would no doubt be reduced. The reply that the level and structure of state programs are the business of voters in each state is not fully adequate. By providing supplementary benefits and higher tax rates on top of the federal floor, states could raise federal costs. If high benefits and tax rates reduce work effort and earnings, the required federal payment would increase.

C. *Social Insurance Reforms*

Finally, of vital importance in designing a sensible income maintenance system is the reform of social insurance programs. Expenditures on programs such as social security, medicare, and unemployment insurance, under which people gain rights to benefits based largely on prior work experience, have grown enormously. But they are becoming increasingly inefficient in helping the lowest income groups, and certain other inequities in the system are becoming increasingly widespread. Moreover, increases in social insurance benefits may simply replace money that could have been diverted to other programs.

The best illustration of these problems is the interaction between retirement benefits under social security (*OAI*) and welfare benefits to aged persons to be provided under the new Supplemental Security Income (*SSI*) program which went into operation January 1974. *SSI* now guarantees a minimum income to all aged persons of $140 for a single individual and $210 for a couple. Many states supplement these guarantees, thereby raising income floors for many aged persons. The difficulty comes when one considers *OAI* and *SSI* in combination. *SSI* is a type of negative income tax, with a 50 percent tax rate on earnings over $85 per month, and a 100 percent tax rate on more than $20 of unearned income, such as social security or pension income. This *SSI* policy means: (a) that social security benefits will be worth only $20 per month to the 20 percent of aged social security recipients also receiving *SSI* benefits and to the even higher percentage of *OAI* beneficiaries receiving state supplements; (b) that across-the-board increases in social security benefits will provide no help to the lowest-income aged; (c) that *SSI* will have a leveling effect by giving equal benefits to retired workers with vastly different covered earnings; and (d) that availability of *SSI* coverage for the needy at age 65 will eliminate for many people the potential loss from reduced social security benefits normally borne by those who retire early.

Increasingly, benefits from welfare elements in social security payment formulas will go to middle- and upper-income aged, often those who have other public and private retirement income. However, low- and moderate-income workers will bear a large part of the costs of these welfare elements, since they tend to shift the tax burden from the personal income tax toward the social security tax.

Unless there is improved program coordination between *SSI* and social security and restructuring of social security along stricter pension lines, social security will become a poorer buy for more workers, and higher expenditures on social security will waste scarce resources in a futile and inefficient effort to help the poor.

IV. Where Do We Go From Here?

We believe that progress toward reform depends on several factors: whether we can hold down the number of new programs with benefits which decline with income; whether we can bring benefits for male-headed families closer to those for female-headed families; whether we can conserve resources by separating the welfare and pension elements in social security; whether we can develop better information on the distribution of current benefits; and, finally, whether we can waive the "no-loss" principle for current recipients so as to achieve gains in system rationality. Most of what has happened in the last five years does not make us sanguine about these prospects. On the other hand, both the Congress and the Administration are devoting time and money to studying the income transfer system as a whole and *that* is history making.

REFERENCES

M. Feldstein, "Unemployment Compensation: Adverse Incentives and Distributional Anomalies," Harvard Instit. of Econ. Res., disc. pap. 317, 1973.

W. A. Klein, "Problems in Choosing Family Unit Rules for a Negative Income Tax," Instit. for Research on Poverty, disc. pap. 61–70, Madison 1970.

I. Lurie, "Current Public Assistance Benefits and An Assessment of State Supplementation Under Proposed Federal Alternatives," *Studies in Public Welfare*, pap. no. 7, Subcomm. on Fis. Pol., Joint Econ. Comm., Washington 1973.

B. Okner, "The Role of Demogrants as an Income Maintenance Alternative," *Studies in Public Welfare*, pap. no. 9, Subcomm. on Fis. Pol., Joint Econ. Comm., Washington 1973.

A. Packer, "Categorical Public Employment Guarantees: A Proposed Solution to the Poverty Problem," *Studies in Public Welfare*, pap. no. 9, Subcomm. on Fis. Pol., Joint Econ. Comm., Washington 1973.

J. Storey, A. Townsend, and I. Cox, "How Public Welfare Benefits Are Distributed in Low-Income Areas," *Studies in Public Welfare*, pap. no. 6, Subcomm. on Fis. Pol., Joint Econ. Comm., Washington 1973.

U.S. Budget for Fiscal Years 1969 and 1974, Washington.

U.S. Dept. of Housing and Urban Development, *Housing in the Seventies*, Washington 1973.

THE PROPERTY TAX: PROGRESSIVE OR REGRESSIVE?

A New View of Property Tax Incidence

By HENRY AARON*

Many economists are fond of quoting Keynes's peroration in the *General Theory* on the unwitting use by politicians of ideas put forth by economists a generation earlier. It makes us feel ahead of our times. The truth may be less flattering, that economic research comes too late to elevate popular debate in the current crisis and a generation ahead of the next. In the hope that this is not the case, I put forth three propositions in this paper, which challenge the conventional view of the property tax:

(1) While the most widely quoted studies support the view that the property tax is highly regressive, this work rests on incidence assumptions unsupported by any plausible theory. Indeed these empirical estimates are the product of rules of thumb that grow more implausible the more closely one examines them.

(2) A more refined empirical application of these flawed rules of thumb would lead to the conclusion that the property tax is far less regressive than the most widely quoted evidence suggests and may well be progressive.

(3) Empirical estimates of property tax incidence incorporating relevant theoretical advances would show the property tax to be progressive; but such estimates are not likely to be made in the near future because of the lack of data.

I.

Nearly all popular discussions of property taxation take its regressivity for granted.[1] This assumption is supported by official publications of the *U.S.* government, e.g., U.S. Bureau of the Census (1973a); quasi-governmental organizations such as The Advisory Commission on Intergovernmental Relations (1973); research centers (the Tax Foundation); and many academic economists.[2]

Most such studies assume that (1) homeowners bear property tax burdens directly in their capacity as occupants and are unable to shift the tax to anyone else; (2) owners of rental properties shift the tax substantially to renters who bear the tax in proportion to rents paid; and (3) owners of other taxed property pass on to final consumers a sizeable fraction (half

---

* Economic Studies Program, the Brookings Institution, and Economics Department, University of Maryland. This paper is a progress report on work I am doing at Brookings which will appraise present views of the differential incidence of the property tax in the light of recent economic analysis. The views expressed are my own and not necessarily those of the officers, trustees, or other staff members of the Brookings Institution or the University of Maryland. I wish to thank William Perkins and Philip Spevak for research assistance, and John Brittain, Frank de Leeuw, Robert Hartman, Benjamin Okner, Joseph Pechman, and Robert Reischauer for comments on an earlier draft.

[1] Throughout this paper, the framework of differential tax incidence is employed. This approach is relevant to debates about the desirability of reducing the degree to which a given amount of public expenditures should be financed from property taxes. It is not relevant to other important questions concerning the impact of changes in the public expenditures and property taxes, other revenues held constant.

[2] See W. I. Gillespie; Richard A. Musgrave, Karl E. Case and Herman Leonard; R. Musgrave and Peggy Musgrave; Dick Netzer (1966); Horst Recktenwald; and Carl Shoup.

according to the Musgraves, variable fractions, depending on the type of property, according to Netzer) of nonresidential property taxes. Presumably because of inadequate data, these studies do not distinguish between land and improvements in allocating burdens or take account of the very wide variations in effective tax rates across and within jurisdictions.

The chief question about these studies concerns the theoretical justification for the rules of thumb used in allocating tax burdens. What model yields the conclusion that property taxes will be borne almost exclusively by users of commodities or services produced by taxed capital? This view might be justified if the tax raised the price of housing by the amount of property tax, leaving the net-of-tax price unchanged. For the price of commodities produced with taxed capital to rise by the amount of the tax, the net-of-tax price of capital must be unaffected by the property tax. Such stability is possible if owners of capital can and do raise the price of commodities by varying proportions depending on the ratio of property taxes to total cost, and if the resulting shifts in demand by consumers in response to changing relative prices are so trivial that movements of capital among industries may be safely ignored.[3] Since the property tax, viewed as an excise, varies from perhaps 30 percent or more on housing in high tax jurisdictions to virtually nothing on some services, this view of the property tax demands that one ignore the basic proposition of demand theory that quantities demanded depend on relative prices if one is to accept the cited allocations. Similarly, one would have to ignore the possibility of substitution between land and capital, on the one hand, and current inputs, on the other, identified by Richard Muth. Moreover, since property tax rates would alter relative prices among taxing jurisdictions, one would have to assume that movements of labor among labor markets with different tax rates are unaffected by relative real wages or that any induced movement of labor leaves relative factor prices unaffected. Altogether these theoretical implications of the rules of thumb used in the cited studies seem so implausible that one is driven to seek some other model.[4]

The property tax also might be borne in proportion to consumption of housing and of other goods if it were the tax price for public services levied in proportion to such consumption. Indeed, Bruce Hamilton has shown that under rather stringent assumptions the property tax is precisely a price charged for local public goods received.[5] Unfortunately, this approach—one of balanced budget incidence in which the burden of the property tax and the value of local public goods paid for by the property tax exactly offset one another—is quite out of spirit with the studies cited. Furthermore, however well Hamilton's approach fits the facts of suburban zoning and finance, it is not well suited to an analysis of the distributional consequences of zoning and taxation in large cities.

---

[3] The net-of-tax price of capital would also be unaffected if the supply of capital were infinitely elastic at the existing rate of return, a condition I regard as absurd and ignore henceforth.

[4] Note that these strictures do not apply fully if one is examining the distributional effects in one community of a change in property tax rates in that community. In that case the price of capital to the community changes negligibly, but if one ignores any induced change in the technology of producing housing services, one would expect declines in land values (or in wages, if labor is a quasi-fixed factor) to offset exactly the tax differential. The price of those commodities sold outside the community would not rise unless the local producer was the principal national supplier; property taxes levied on capital used in the production of exported goods sold in competitive markets would be absorbed by owners of capital, by owners of land through decreased rents or land values, or by labor through lower wages. Since the studies cited deal with the distribution of burdens of the entire property tax and not of differentials in local rates, however, the strictures apply with full force.

[5] See Hamilton and a more refined version in Hamilton, Craig MacFarlane and Peter Mieszkowski. The property tax becomes a price when zoning restrictions operate to create communities homogeneous with respect to housing and public good consumption.

I conclude that the property tax should not be thought of as regressive unless and until a model can be found to justify this view.

II.

In addition to lacking theoretical foundations the notion of regressivity is subject to empirical challenge on at least two grounds.

First, calculations alleging regressivity usually classify households on the basis of annual measured incomes which imperfectly reflect long-run economic circumstances. Since housing expenditures involve long-term commitments unlikely to be adjusted fully for short-run fluctuations, classification of households according to current measured incomes biases estimates of the distribution property tax burdens in the direction of regressivity for precisely those reasons which have led students of consumption behavior to conclude that the ratio of normal ("permanent") to measured income diminishes as one moves up the scale of measured income. We ordinarily classify households with respect to current measured income for the allocation of tax burdens and disregard any biases which this procedure may entail. (For an exception, see Thomas Mayer.) With respect to personal income taxes this procedure causes only minor distortions in estimates of tax burdens by income class, since apart from loss carry-back and carry-forwards and deductions related to normal income, the tax liability of a household temporarily at one income level is no different from that of a household normally at that income level. While the use of measured income as a proxy for normal income does not seriously distort inferences about the income tax rate at any level of normal income, it does distort inferences about the size distribution of normal income and hence the long-run distribution of tax burdens. It also seriously distorts empirical measures of the burden distribution of taxes related to normal income.

If the residential property tax is distributed in proportion to housing consumption, then, subject to a qualification to be presented below, it is progressive, proportional, or regressive with respect to normal income as the normal income elasticity of demand for housing is greater than, equal to, or less than one. The evidence on housing demand is mixed—see Aaron, Geoffrey Carliner, de Leeuw, Sherman J. Maisel, James B. Burnham, and John S. Austin—but an income elasticity near one seems reasonable on the basis of currently available research.

Even if the elasticity of housing expenditures with respect to normal income is below one, the portion of the property tax that falls on rental housing may well be progressive. The property tax is levied on capital value, not rents. As George Peterson has pointed out, the ratio of market value to rent increases with rent. This phenomenon occurs because inexpensive rental housing typically is located in older buildings with relatively short expected lives in which maintenance expenditures constitute a relatively large fraction of gross rents and because owners of relatively expensive housing may get a larger part of their income through capital gains. This effect is seriously undercut by administrative procedures which cause assessment ratios on inexpensive housing to exceed those on expensive housing; see U.S. Bureau of the Census (1973b, p. 15). It is impossible with presently available data to determine the balance between these two opposing forces. However, if reform of property tax administration should result in equal assessment ratios, it is likely that the property tax on rental property would be converted into a proportional, and possibly progressive, tax even if it is viewed as an excise on housing.

Table 1 indicates that even if the property tax on housing were viewed as an excise, it would be progressive if it were well-administered. It shows income elasticities of rental expenditures, home value, tax rate and property tax payments estimated on the assumptions that the property tax is shifted forward in full to renters and homeowners (in their capacity as consumers) and that the tax is administered with complete equity (a condition clearly contrary to fact). The estimates suggest that under these circumstances the property tax on homeowners would be mildly, but significantly, progressive—with an income-elasticity of approximately 1.2—while that on renters would be roughly proportional, with an elasticity between 0.8 and 1.1. (See Table 1; note that for examining global incidence the simple regressions are probably more instructive than the multiple regressions.)

III.

As indicated in Section I, however, the property tax cannot reasonably be regarded as an excise on housing services and other commodities produced with taxed capital. Building on Arnold Harberger's classic analysis of the corporation income tax, Peter Mieszkowski (1967, 1969, 1970, 1972) has produced a coherent base for differential tax analysis of property tax incidence. While enormous simplifications and qualifications are necessary in applying Mieszkowski's analysis to empirical work, economists now have a better theoretical foundation than ever before for estimates of property tax incidence.[6] Space

TABLE 1—ESTIMATED INCOME ELASTICITIES[a]

Group and Variable	Simple Regression[b]	Multiple Regression[c]
Renter		
Rent	0.561 (10.4)	0.336 (5.3)
Tax rate	0.008 (0.1)	−0.004 (0.1)
Property tax (unadjusted)[d]	0.569 (6.4)	0.491 (5.6)
Property tax (adjusted)[d]	0.797–1.138	0.687–0.982
Homeowners		
Home value	1.132 (13.7)	0.995 (10.4)
Tax rate	0.114 (1.6)	0.013 (0.2)
Property tax	1.246 (10.8)	1.077 (10.4)

[a] Income, Y, is defined as a weighted average of annual incomes (with declining weights) for 1967–71, i.e., as

$$\left[\sum_{-4}^{0} Y_i(1+0.2i)\right]\Big/3$$

where Y_i is total family real net income plus estimated windfall income in the year i, obtained from the Panel Study.

[b] Simple regression results are the coefficient b and (in parentheses) its t value, obtained from the regression $\log r=a+b \log Y$, where r is the relevant variable for 1971.

[c] Multiple regression results are the coefficient b and (in parentheses) its t value, obtained from the regression $\log r=a+b \log Y+\sum c_i \log z_i$, where the z_is include dummy variables with significant coefficients for size of city, distance from city of 50,000 or more, Census region, marital status and number of children, positive savings equal to less than two months of income, savings equal to more than two months of income, business ownership and race.

[d] The income-elasticity of property tax payments (unadjusted) is calculated on the assumption that property value is 100 times monthly rent. The adjusted estimates take account of George Peterson's finding that the elasticity of value with respect to rent is above 1 (1.4 to 2.0 in his data), and equal the unadjusted estimates multiplied by 1.4 or 2.0.

Source: Sample of households that moved within the past year (renters) or two (homeowners), drawn from the University of Michigan's Panel Study of Income Dynamics. Property tax burdens estimated on the assumption that the average property tax rate for cities of 50,000 or more and the rest of county, reported by *U.S.* Bureau of the Census (1973b), applied to each renter and homeowner. Data from Federal Housing Administration reports on effective property tax rates for some counties not covered in the Census report were also used.

---

[6] Whether better is good enough is open to dispute. In forsaking his previous estimates, Netzer (forthcoming) argues that the property tax is just too complicated to lend itself to global incidence calculations. It may be that Netzer is challenging the usefulness of all estimates of global tax incidence, since the grounds for awarding the property tax first place in the complexity sweepstakes, ahead of the corporation income tax and state and local income taxes, are far from clear.

constraints prohibit more than a stark outline of this analysis, which is already entering textbooks and popular writings, e.g., George Break, Musgrave and Musgrave (1973) and Recktenwald.

If factor supplies are fixed in amount and perfectly mobile, and markets are competitive, a uniform tax on capital will be borne completely by owners of capital. But, the property tax does not apply equally to all capital goods. First, a large part of real capital is wholly exempt, including all publicly owned capital, a large fraction of business inventories and machines, real estate owned by nonprofit enterprises, and some exempted private real estate. Second, average effective rates of property tax vary widely across jurisdictions. Third, within jurisdictions effective rates vary widely among and within classes of taxable property, occasionally with legal sanction, often because of extralegal decisions by assessors, and frequently because of administrative inefficiency or lack of resources. As a result, a wide distribution of effective rates applies to capital goods, ranging from zero on exempt capital to 5 or 6 percent of value per year and even more in some cases. Capital subject to above- (below-) average tax rates would have to generate above- (below-) average gross yields to leave a net yield as high as could be earned elsewhere. Whether capital bears a burden equal to, greater than, or less than the tax depends on relative factor intensities, elasticities of substitution between factors, and elasticities of demand.

We next note that in an analysis of property taxation it is not legitimate to assume perfect factor mobility. Part of the tax may fall on land. Differentially higher property taxes will cause land values and rents to decline if the supply of land is fixed. Alternatively if the "world" in which the tax applies includes only a small part of available land, and if land can be converted at a constant cost to or from some "backstop" use, such as agriculture, then differences in property tax rates will affect the size of communities, but not the price of land; see Mieszkowski (1972).[7] If labor is immobile, then capital in high tax communities may escape at least part of the burden of the property tax through (a) increases in prices of locally produced goods not subject to competition from commodities produced in jurisdictions with lower taxes or (b) reductions in wage rates paid to workers employed in producing commodities the price of which is determined in markets extending beyond the community. In low tax communities, one would find symmetric commodity, wage and rent "subsidies."

Thus we do not assume perfect mobility or tax uniformity. But the following analysis embodies two other assumptions that should be made explicit. The first is that price elasticities of supply for both labor and capital are zero. Little has been written about the case in which they are not, but, as Martin Feldstein (1972) has shown, if the elasticity of labor supply is not zero, a reduction in real wages due to the property tax would have a negligible long-term impact on income distribution because the supply of labor is dominated in the long run by population growth and because the capital-labor ratio will adjust for any change in the labor force to its long-term growth path. (However, the transitional effects may be quite protracted.) If the property tax depresses the rent-wage ratio, the outcome then depends on savings propensities. Analyzing a general tax on capital, Feldstein (1973) presents numerical examples suggesting that if savings fall due to the tax, most of the tax burden—one-half to two-thirds—is still borne by capital.

---

[7] Under these circumstances, it should be noted, the site value tax, commonly defended for its neutrality, would not fall solely on rents and would affect the allocation of real resources. See Mieszkowski (1970).

Second, the analysis assumes either that owners of capital maximize profits or that the extent of their failure to do so is not correlated with changes in property tax rates. The incidence of the property tax thus consists of three effects: (a) a capital tax effect (the burden of the average tax rate on capital is distributed in proportion to the ownership of capital); (b) an excise tax effect (deviations around the average tax rate cause shifts in capital among regions or industries, possibly raising or lowering the gross yield of capital and the share of output accruing to labor); and (c) an immobility effect (deviations of tax rates around the average would impose burdens or generate gains in a complex fashion, raising or lowering prices of local goods not subject to external competition, such as housing, some local wages, and some rents).

Thus, an attempt to allocate empirically the property tax burden would, following the three effects just listed, proceed in three steps. In the first, the average effective rate of property taxation would be allocated among all owners of capital. The second step requires an estimate of how tax-induced movements of capital among activities and regions have affected the average before-tax yield of capital relative to the wage rate. Housing services are produced by a more capital-intensive technology than that used in the production of any other commodity or service (Edward F. Denison, p. 124); property taxes levied on residential structures are higher than the average rate calculated on all capital; see Netzer (forthcoming, Table 4); and residential real estate includes nearly half of all capital subject to property taxation; see Netzer (forthcoming, Table 3). In this situation the differentially higher tax on capital in the relatively capital-intensive housing industry will cause that industry to release capital and labor in proportions that can be absorbed by other industries only at a decreased ratio of before-tax profits to wages.[8] Finally, a complete analysis of property tax incidence should include a distribution of the excise-type effects. Empirically, this task is much the hardest. In general, prices of commodities not subject to external competition will tend to be raised (or lowered) by the property tax in those communities with tax rates above (or below) average. In addition one would need to determine whether above- or below-average property taxes tended to depress or inflate wages, profits, or rents on those commodities subject to competition from external markets.

While lack of data makes work on some elements of this agenda impracticable, estimates of the most important magnitudes for approximations of global incidence can be obtained. A distribution of the average rate of property taxation by net worth is shown in Table 2. As is apparent, the distribution is "J"-shaped when households are classified by annual income. This pattern arises, despite the fact that nonhuman capital is distributed progressively with respect to permanent income, because households with negative transitory income and the aged are concentrated in low income brackets.

The effect on the rate of return of shifts in capital from high-tax to low-tax sectors has not been estimated, but I conjecture that when and if such work is done it will indicate that the relative return to capital is reduced (see footnote 8).

---

[8] An illustrative application of the basic Harberger-Mieszkowski equation makes this point clear (see Harberger). Assume that the property tax is applied at a 2 percent higher rate to an industry that employs half of all capital with a capital-labor ratio of two than to another industry employing the other half of capital with a capital-labor ratio of one-half. Further assume unitary elasticities of substitution and supply. In that case the price of capital would be depressed by 1 percent, half of the tax differential. Unfortunately such examples can only be illustrative, since empirical measures of relevant elasticities are unavailable.

TABLE 2—PROPERTY TAX RATE BY INCOME CLASS, DISTRIBUTION IN PROPORTION TO OWNERSHIP OF CAPITAL, 1966

Income Bracket[a]	Property Tax Rate as Percent of Income
$ 0–$ 3,000	2.5
3,000– 5,000	2.7
5,000– 10,000	2.0
10,000– 15,000	1.7
15,000– 20,000	2.1
20,000– 25,000	2.6
25,000– 30,000	3.8
30,000– 50,000	4.4
50,000– 100,000	6.2
100,000– 500,000	8.2
500,000–1,000,000	9.6
1,000,000–and over	10.1
Average	3.0

[a] Income equals adjusted gross income, transfer payments, tax exempt interest, and excluded realized long-term capital gains.
Source: Joseph Pechman and Benjamin Okner.

The empirical distribution of the excise tax effects has also not yet been estimated; to do so will be quite difficult. First, what should be done about the roughly 22 percent of physical capital owned by governments? As Netzer (forthcoming) observes, much public borrowing— e.g., that by local governments and special districts—is sensitive to the cost of capital. To the extent that such exemptions are accorded by local governments to their own real estate, we may regard them merely as a bookkeeping device, through which governments avoid the charade of transferring resources from themselves as capital users to themselves as tax collectors. Such an exemption does not give rise to excise effects. No one has shown that demands for capital by the federal government are interest sensitive. The imputation of excise tax effects enjoyed by tax exempt private organizations is probably impossible. On balance, I conclude that publicly owned and exempt private capital should be excluded from analysis of property tax incidence, although it might be argued that some rule of thumb should be used to impute ownership of such capital to private households.

Excise effects also arise from variations in tax rates among jurisdictions or classes of property. To the extent that geographical variations in rates are capitalized in land rents, that portion of the tax is distributed according to ownership of land at the time the tax is levied or increased. The distribution of property tax burdens capitalized into land values would require that the investigator trace ownership from the time the tax was imposed or increased through subsequent heirs to current owners. Since land ownership is distributed very progressively with respect to income (see Mason Gaffney), such capitalization appears to entail redistribution among high income households in various parts of the country.

To the extent that variations in property tax rates cause variations in prices of goods not subject to external competition or cause variations in wages, one would need some indication of the relationship between property tax rates and income by jurisdiction.[9] In regressions not reported here I found that high (low) per capita income and high (low) property tax rates are associated strongly among states; weakly across counties in five states for which I have obtained data; and strongly across jurisdictions with populations of 50,000 or more whose effective property tax rates are reported by the Census. These relationships describe nothing more profound than the fact that both tax rates and incomes tend to be low in the south and in rural areas. On the other hand, within samples for each of seven urban counties in New Jersey, high income communities tended to have lower property tax rates than did low income communities; see also Advisory Committee on

---

[9] Since excise tax effects alter commodity prices, it is the income of households residing in the jurisdiction, not that of property owners, that is relevant.

Intergovernmental Relations (1967). Since labor mobility within counties is likely to be greater than that among counties, states, or cities of 50,000 or more, it is likely that a greater part of intracounty tax differentials are capitalized directly into land price differentials and are borne by land owners (see Wallace Oates, A. Polinsky and L. Rubinfeld, and R. Stafford Smith). These results suggest caution in assuming that consideration of excise tax effects restores the conclusion that the property tax is regressive. The opposite conclusion seems more probable.

One must still ask whether consumption of heavily (lightly) taxed commodities such as housing (medical care) is regressively or progressively related to income. Table 3 indicates that the proportion of income devoted to items that are subject to low property tax rates is quite regressive over measured income classes though not so regressive as housing. As a fraction of current consumption (a proxy for permanent income if the permanent income elasticity of consumption is unitary), these items are roughly proportional across measured income classes.

On balance it appears that the property tax is distributed with respect to measured income not very differently from the pattern shown in Table 2. With respect to income measured over a longer period, the property tax is probably progressive in the aggregate. While these statements are subject to enormous qualification and require more empirical support, it would seem that the property tax should be presumed progressive unless proved regressive, instead of the reverse presumption embedded in the conventional wisdom.

IV.

The foregoing analysis is relevant to at least two significant public policy questions. First, should the average level of property taxation be reduced or increased? Many considerations other than tax incidence are relevant to this question. As regards incidence, one's position should depend on one's views both on the desirability of progressive taxation and on alternatives to the property tax. The major conclusion

TABLE 3—DISTRIBUTION OF SELECTED CONSUMER EXPENDITURES AS PERCENT OF INCOME AND CONSUMPTION, BY INCOME BRACKET[a]

	Percent of Income Devoted to		Percent of Consumption Devoted to	
Income Bracket	Shelter	"Low-Taxed Consumption"[b]	Shelter	"Low-Taxed Consumption"[b]
Under-$1,000	39.2	37.4	17.6	16.8
$1,000- 1,999	19.7	17.3	17.1	15.0
2,000- 2,999	15.6	15.7	15.3	15.4
3,000- 3,999	13.5	14.8	13.9	15.2
4,000- 4,999	11.8	13.0	13.1	14.5
5,000- 5,999	11.3	12.8	13.2	15.0
6,000- 7,499	10.5	12.6	12.9	15.4
7,500- 9,999	9.3	11.9	12.2	15.6
10,000-14,999	8.1	11.7	11.6	16.7
15,000 and over	6.4	8.7	12.5	17.0

[a] Survey of Consumer Expenditures 1960–61.

[b] "Low-taxed consumption" includes the following consumption categories: personal care, medical care, recreation, reading, education.

of this paper is that one cannot legitimately adduce regressivity of the property tax as a reason for cutting back *general* reliance upon it.

Second, should a particular jurisdiction cut back or increase reliance on the property tax? This question involves an examination only of the excise effects of changes in a community's property tax rate relative to the national average. A large part of any change in property taxes implemented in a jurisdiction that is a small part of a labor market area is likely to be capitalized in changed property values. If an entire labor market area, say a standard metropolitan statistical area, could accomplish such a shift, it would change the cost of capital to the community and might affect wages as well as property values. The incidence of such a change is unclear.

The foregoing remarks are not intended as a brief for greater reliance on property taxes by states or local governments. Fiscal decisions should hinge on considerations in addition to tax incidence, including perceptions of ability to pay, and benefits received, convenience of collection, and impacts on migration of population and industry. While the trend away from property taxes may properly continue, it should not be justified by the view that its burdens are distributed regressively.

REFERENCES

H. Aaron, *Shelter and Subsidies: Who Benefits from Federal Housing Policies?* Washington 1972.

G. Break, "Taxation: Incidence and Economic Effects," mimeo.

G. Carliner, "Income Elasticity of Housing Demand," *Rev. Econ. Statist.*, Nov. 1973, *55*.

F. de Leeuw, "The Demand for Housing: A Review of Cross-Section Evidence," *Rev. Econ. Statist.*, Feb. 1971, *53*, 1–10.

E. Denison, *Why Growth Rates Differ*, Washington 1967.

M. Feldstein, "Incidence of a Capital Income Tax in a Growing Economy with Variable Savings Rates," Harvard Instit. of Econ. Res., disc. paper no. 300, June 1973.

———, "Tax Incidence in a Growing Economy with Variable Factor Supply," Harvard Instit. of Econ. Res., disc. no. 263, Dec. 1972.

M. Gaffney, "What is Property Tax Reform?" *Amer. J. Econ. and Soc.*, Apr. 1972, *31*, 129–53.

W. I. Gillespie, "Effect of Public Expenditures on the Distribution of Income," in R. A. Musgrave, ed., *Essays on Fiscal Federalism*, Washington 1965.

B. Hamilton, "Capitalization of Intra Jurisdictional Differences in Local Tax Prices," mimeo.

———, **C. MacFarlane, and P. Mieszkowski**, "The Efficiency and Distributive Aspects of Zoning," mimeo.

A. Harberger, "Incidence of the Corporation Income Tax," *J. Polit. Econ.*, June 1962, *10*.

M. Krzyzaniak, ed., *Effects of Corporation Income Tax*, Detroit 1966.

S. J. Maisel, J. B. Burnham, and J. S. Austin, "The Demand for Housing: A Comment," *Rev. Econ. Statist.*, Nov. 1971, *53*, 410–13.

T. Mayer, "The Distribution of the Tax Burden and Permanent Income," *Nat. Tax J.*, forthcoming.

P. Mieszkowski, "A Critical Appraisal of Land Value Taxation," mimeo., May 1970.

———, "On the Theory of Tax Incidence," *J. Polit. Econ.*, June 1967, *75*, 250–262.

———, "The Property Tax: an Excise Tax or a Profits Tax?" *J. Pub. Econ.*, April 1972, *1*, 73–96.

———, "Tax Incidence Theory: the Effects of Taxes on the Distribution of Income," *J. Econ. Lit.*, Dec. 1969, *7*, 1103–1122.

R. A. Musgrave, K. E. Case and H. Leonard, *The Distribution of Fiscal Burdens and Benefits*, Harvard Instit. of Econ. Res., disc. paper no. 319.

——— **and P. B. Musgrave**, *Public Finance in Theory and Practice*, New York 1973.

R. Muth, "Capital and Current Expenditures in the Production of Housing," Center for Res. in Econ. Growth, Stanford University memo. 123, October 1971.

D. Netzer, *Economics of the Property Tax*, Washington 1966.

———, "The Incidence of the Property Tax Revisited," *Nat. Tax J.*, forthcoming.

W. Oates, "The Effects of Property Taxes and Local Public Spending on Property Values: An Empirical Study of Tax Capitalization and the Tiebout Hypothesis," *J. Polit. Econ.*, Nov./Dec. 1969, *77*, 957–71.

J. Pechman and B. Okner, "Who Paid The Taxes in 1966?" *Amer. Econ. Rev. Proc.*, May 1974.

G. Peterson, "The Regressivity of the Residential Property Tax," *Urban Instit. Work. Pap. 1207–10*, Nov. 1972.

A. Polinsky and L. Rubenfeld, "The Long-Run Incidence of a Residential Property Tax and Local Public Services," work. pap. 1207–29, The Urban Institute, Washington 1974.

H. C. Recktenwald, *Tax Incidence and Income Distribution: An Introduction*, Detroit, 1971.

C. S. Shoup, *Public Finance*, Aldine 1969.

R. S. Smith, "Property Tax Capitalization in San Francisco," *Nat. Tax. J.*, June 1970, *23*, 177–191.

Advisory Commission on Intergovernmental Relations, *Financing Schools and Property Tax Relief*, A-40, Jan. 1973.

———, *Fiscal Balance in the American Federal System, Vol. 2 Metropolitan Fiscal Disparities*, Washington, Oct. 1967.

Tax Foundation, *Tax Burdens and Benefits of Government Expenditure by Income Class, 1961 and 1965*, New York 1967.

U.S. Bureau of the Census, *1970 Census of Housing, Real Estate Tax Data for Homeowner Properties for the United States and Regions: 1971*, HC(S1), June 1973a.

———, *Taxable Property Values and Assessment—Sales Price Ratios, Part 2, Assessment-Sales Ratio and Tax Rates*, Oct. 1973b.

U.S. Bureau of Labor Statistics, *Consumer Expenditures and Income: Total United States, Urban and Rural 1960–61*, BLS Report No. 237–93, Feb. 1965.

Is a Property Tax on Housing Regressive?

By RICHARD A. MUSGRAVE*

My discussion of property tax incidence will be limited in three ways. First, I shall deal only with that half of the tax which is assessed on housing, this being the part which has received most attention in recent policy discussion. Secondly, I shall address myself to the incidence of a hypothetical national tax. While the local nature of the property tax with its jurisdictional differentials poses important additional problems, these are not central to whether the local property tax as a whole is progressive or regressive. Finally, I shall deal with differential incidence and examine the distributional consequences of substituting a uniform tax on housing for, say, a proportional income tax. Thus, the expenditure side of local finance is held constant and excluded from this analysis.

The regressive or progressive nature of such a substitution needs to be known to reassess the role of the property tax in the overall tax structure. If the burden falls on the consumers of housing services, its incidence tends to be regressive (Musgrave et al., 1951). If, instead, the burden falls on all capital income, incidence tends to be progressive and the property tax should be reclassified accordingly (Arnold Harberger and Peter Mieszkowski). While the former view was traditionally held by economists and still dominates the political forum, there is an increasing tendency among economists toward the latter approach. Hence clarification is needed.

I. Alternative Theories

I begin with the validity of alternative incidence theories or hypotheses.

A. *The Harberger-Mieszkowski Model*

The new (i.e., Harberger-Mieszkowski, hereafter H-M) model postulates the classical assumption of profit maximization and perfect adjustment between various sectors of the capital market. Viewing the incidence of a uniform tax on housing capital in analogy to the corporation tax, imposition of such a tax will lead to an outflow of capital from the housing to the nonhousing sector until net returns are equalized. As a result, such burden as falls on capital is shared equally by capital in all its uses. If the capital stock is fixed and certain conditions regarding production and demand functions are met, gross factor shares of capital and labor will remain unchanged and the entire burden of the tax as viewed from the earnings side will be on capital. Additional distributional changes will result because reallocation of capital from the housing to the nonhousing sector leads to a rise in the price of housing services relative to that of other goods. Thus, there results a so-called "excise effect" which redistributes real income from consumers of housing toward consumers of other products. The net distributional effect of the tax then equals the combined earnings and excise effects.

To move from this theoretical model to the conclusion that the tax is progressive, we need only postulate that the progressive effect on the earnings side (due to a higher ratio of capital to total income at higher levels of income) more than offsets the regressive effect from the uses side (due to a lower ratio of housing to nonhousing expenditures). Since these assumptions seem reasonable, the general interpretation of the H-M model has been that the tax is progressive.

---

* H. H. Burbank Professor of Political Economy, Harvard University.

Certain qualifications may be noted. A first question is whether a distinction should be drawn between land and improvement parts of the tax base. Assuming the total supply of capital to be fixed, both are similar in this respect, but the capital is spatially mobile whereas land is not. This is of obvious importance for the local property tax, but less so for a national tax. Provided that zoning imperfections are ruled out, land no less than capital can be moved from housing into other uses. The burden of the tax on land will thus be spread among land owners in general, be it in residential or other uses. With both capital and land in fixed supply, the analysis of incidence for the two factors (as distinct from the differential local tax) is essentially similar.

A second query relates to the particular nature of housing as a combined consumption and investment good. Whereas choices regarding consumption patterns and the mix of investment portfolios are usually made independently of each other, the decision to hold owner-occupied housing involves both sets of considerations. Consumers with a high consumption preference for owner-occupied (as against rental) housing derive an ownership rent and will hence place a larger part of their assets into such housing, while holding less of other assets. If the tax burden on all capital income is equalized, this does not affect their position from the earnings side; but since their housing expenditures are higher, they suffer a larger excise burden from the income uses side. This may well be a major factor contributing to the unpopularity of this tax.

The joint nature of housing as a consumption and investment good also bears on whether the tax should be thought of as a tax on housing consumption or investment. To simplify matters, suppose that housing consumption involves only capital with no maintenance costs. The tax on housing consumption would then be the same as a corresponding tax on housing property. But if the tax may be looked upon either way, which incidence pattern (consumption or property) is to be applied? The answer is either. In a situation where consumption involves the use of a capital resource only, the incidence of a consumption tax is in fact the same as that of a tax on the capital used to provide that consumption service. In this sense, the property tax may be viewed as a tax on housing consumption, but under the assumptions of the H-M model, the incidence conclusion is not changed thereby. Within the framework of the H-M model, incidence continues to be in line with the distribution of capital income, supplemented by the excise effect from the uses side.

Finally, the result may change if the assumption of variable factor supplies is introduced. Long-run incidence in this case may impose a burden on labor as well as capital, so that the progressivity of the burden distribution (not only for the property but for all progressive forms of tax) tends to be reduced. (See Martin Feldstein, Marion Krzyzaniak, and Kazuo Sato.) However, such effects take a long time to work themselves out and other policy instruments (e.g., the state of budgetary balance) may be brought into the picture to maintain the savings and growth rate.

B. *Administered Pricing Models*

While recognizing the admirable contribution to the theory of incidence rendered by Harberger's analysis of the corporation tax and Mieszkowski's application thereof to the property tax, I must insist that the logical consistency or beauty of such models is no proof that they describe what in fact happens. The assumptions of profit maximization and perfect capital mobility are useful, and the economist's task would be simplified if the world in fact met these assumptions. But the Creator's concern with the happiness of economists might not have sufficed to arrange things this way. Seeing the working of the markets as I do, I must allow for the implications of alternative assumptions. The problem then becomes one of empirical evidence and not merely of logical consistency.

While such evidence on property tax incidence is scanty, some general observations may be made. Regarding the part of the tax on owner-occupied residences, I see little reason to expect that purchasers of houses possess unused market power which they release in response to the tax. Though the previously noted aspect of ownership rent becomes of increased importance—housing, for instance, provides otherwise unavailable leverage for inflation hedging—the competitive model is appropriate on the earnings side.

I feel less comfortable with this model for rental housing. While comprising only 20 percent of the total housing base, rental housing is of special concern since it weighs heavily at the lower end of the income scale. Housing markets, especially for low-income housing in central city settings, may well be markets where landlords act as restrained monopolists (being threatened by rent control and/or public rage) who find it justified to raise rents as taxes are increased, or where tax increases may act as a signal to landlord-oligopolists to raise rents. Rate reductions in turn may or may not have opposite effects. While the fact that rental contracts or rent controls frequently allow for tax increases is no proof that the tax is passed on to consumers through the mechanism of imperfect markets, it at least raises such suspicion. Given inelastic demand for low cost housing and a relatively elastic short-run supply through varying levels of maintenance, these possibilities cannot be ruled out. I would suspect that a substantial amount of initial shifting into higher rentals does in fact result. This being the case, I do not accept the extreme assumption that no part of the tax is shifted to the tenants. Rather, I prefer to allocate a substantial part of the tax on this basis.

Consumer burdens resulting from *this* type of shifting, moreover, must *not* be confused with the excise burden of the H-M model. The latter operates as a consequence of the capital flow induced by the initial burdening of capital income in the taxed sector and involves a redistribution between consumers of housing services and of other products. The burden-shifting now visualized involves an initial passing on of the burden from landlords (on whom the statutory burden rests) to tenants. Landlords effect this by drawing on previously unrealized monopoly power, thus leaving capital income initially free of burden (or in any case with a lesser burden) and involving a burdening of tenants rather than a redistribution among consumers. While further shifts in capital allocation and excise effects will follow, the end result will differ (i.e., leave a heavier burden on tenants and be more regressive) from that under the H-M model.

The decisive difference between this and the H-M model, therefore, is not between partial and general equilibrium analysis, but between models that do and others that do not require profit-maximizing behavior and perfect capital mobility. A theory which allows for imperfections and unconventional firm behavior, I insist, can be the first move in a general equilibrium analysis no less than one which rules them out. The outcome of the model depends on its behavioral assumptions, including those which determine the initial response to the tax. It is, after all, but a dispensing machine, where the juice you draw depends on which button you push. While model builders like to assume away imperfections and market behavior not aimed at profit maximization, it does not follow that no such things exist. It just means that the economist's job is more difficult and that an understanding of institutions becomes more important.

C. *Addendum*

Thus far the discussion has concerned a nationwide and uniform tax on housing. Before proceeding, I note briefly what will happen if other features of the existing property tax are allowed for.

Local Tax

The property tax is not nationwide but differs by jurisdictions. While the average effective rate of tax may be treated in terms of a national tax, departures therefrom exert a second type of excise effect. Generated by the movement of capital toward low rate jurisdictions, this is the excise effect emphasized in Mieszkowski's formulation. On the earnings side of incidence, the net rent of land and the wages of immobile labor will fall in high and rise in low tax jurisdictions. On the uses side, the excise effect burdens immobile consumers of housing services in high tax jurisdictions. While it is reasonable to expect these interjurisdictional excise effects to be regressive, they are not likely to change the general conclusion reached for a national tax. All this, of course, is compatible with the possibility that the incidence of a local tax increase, taken by itself, will differ from that of a national tax or, for that matter, of an increase in the average rate of local taxes.

Tax on Nonhousing Property

The property tax is not a tax on housing only. Nearly one-half of the revenue is collected from business property other than rental. As a result, the interindustry excise effect becomes less significant. However, the question of market structure now arises not only in the rental market but in other industries as well. While I retain my suspicion that a substantial part of the corporation tax is shifted, I find less basis for this with regard to the business part of the property tax. Property tax revenue tends to be derived less heavily from large corporations which are more likely to possess powers of administrative pricing. This suggests a lesser shifting rate. Accordingly, I would not now assume that the tax is fully passed on to consumers, as we did in our estimates of some twenty years ago (Musgrave et al. 1951). At the same time, circumstances which might lead the corporation tax to be shifted (in my sense of administered price adjustments) cannot be excluded from this case. I therefore continue to consider some allowance for shifting—involving, say, one-third of the tax on nonhousing property—to be a less extreme and more plausible assumption than the zero rate implied in the H-M model. As a result, I would expect the tax on nonhousing property to be less progressive than that on owner-occupied residence, though more progressive than that on rental housing.

Assessment Differentials

While the actual tax is more general than a tax on housing, it is by no means as general (even within any one jurisdiction) as would be a uniform tax on all property. Business property on the whole tends to be taxed more heavily than residential property; and low-income rental housing may be assessed more heavily (relative to property value, though not necessarily relative to rental income) than are higher income residences. Such differentials further qualify the incidence conclusions with the net effect on progressivity unclear in its direction.

II. Patterns of Incidence under Alternative Theories

I now turn to the alternative patterns of burden distribution which result under the various hypotheses.

A. *Comparative Patterns: Cross-Section Data*

Table 1 shows such burden distributions based on annual cross-section data. Allocation by capital income (Column II) gives the most progressive result, but even here the property tax is by no means progressive throughout. In fact, it is progressive for only the top 20 percent of households and regressive at the lower end. In Column III, where a hypothetical excise effect is allowed for, the falling segment of the U-shape reaches higher, but the change is not significant except for the bottom of the scale.[1] While the total amount of excise may well be substantial (taken here at $10 billion), the distributive effects tend to wash out because the pattern of burden distribution (in line with expenditures on housing services) is largely offset by that of gain distribution (in line with other consumer expenditures). If we use a mixed pattern where the tax on owner-occupied residences is allocated by all capital income while that on rental housing is imputed to tenants (Column IV), the distribution becomes regressive through 95 percent of the scale, with only the upper 5 percent progressive. Full allocation by housing consumption (Column V) renders the tax regressive throughout the scale.

The evidence is that under almost any assumption, including Column II, the tax on housing property remains regressive over the lower to middle end of the income scale. This differs sharply from the effective rate pattern of an equal yield-income tax (see Column VI), which is distinctly progressive over this range. To put the matter differently, it may be noted that under both taxes the top quartile of taxpayers bears about two-thirds of the burden. The respective property tax shares for the first, second, and third quartiles are 4, 12, and 17 percent for the property tax (allocated by capital income) as against 2, 8, and 23 percent for the income tax. To the extent that concern with progression is pointed toward the lower-middle rather than the upper range, this hardly makes the housing tax a desirable form of taxation. As far as the lower end of the scale is concerned, the crucial distinction between taxes lies not so much in the particular base which is chosen (i.e., capital versus

---

[1] The excise effect (equal to the difference between Columns II and III) is computed as the difference between the burden component obtained by allocating $10 billion in line with housing expenditures and the benefit component obtained by allocating the same amount in line with nonhousing consumption expenditures. This $10 billion reflects the estimated increase in the cost of housing outlays due to the price rise of housing services. This amount is based on the following procedure and certain simplifying assumptions regarding production and demand functions (Harberger, p. 219):

a. If we take the observed value of the housing stock after imposition of the tax to be $750 billion and assume a net rate of return of 6 percent, the net return is $45 billion. Adding the yield of a housing tax of $15 billion, we obtain a gross return of $60 billion. Relating this to the housing stock of $750 billion, we get a gross yield on housing capital of 8 percent.

b. If the value of nonhousing capital is estimated at $1,000 billion, the value of the total capital stock is $1,750 billion. Applying the tax yield of $15 billion to this total, we get a tax rate of .86 percent. If we assume again a net yield of 6 percent, the gross yield on total capital is 6.86 percent.

c. The housing excise therefore equals 8–6.86 = 1.14 percent.

d. Under the assumption of unit elasticity of substitution between capital and labor in both the housing and the nonhousing sectors, and of unit elasticity of demand, the increase in housing cost may be estimated by applying the 1.14 percent excise rate to the housing stock of $750 billion, given an excise cost to consumers of housing services of $8.5 billion.

e. Alternatively, the 1.14 percent may be applied to the value of the housing stock, K, as it would have been in the absence of tax. Computed as $750 \times .08 = K \times .0686$, the housing base is now $874 billion and the excise burden becomes 1.14 percent thereof or $10 billion. The illustration of Table 1 uses the upper limit figure of $10 billion.

TABLE 1—INCIDENCE OF PROPERTY TAX ON HOUSING, 1968 LEVELS[a]

		Tax as Percent of Income				
		Property Tax on Housing Allocated by[c]				
Income Brackets[b]	Percent of Families I	All Capital Income[d] II	All Capital Income with Excise Effect[e] III	Mixed[f] IV	Housing Expenditures[g] V	Equal Yield Income Tax[h] VI
0- 4,000	19.9	1.2	1.7	3.2	4.1	0.4
4,000- 5,700	9.6	2.0	2.2	3.0	3.0	0.5
5,700- 7,900	11.5	1.7	1.7	2.3	2.5	1.1
7,900-10,400	12.4	1.1	1.1	1.7	2.3	1.3
10,400-12,500	12.9	0.9	0.8	1.2	2.2	1.4
12,500-17,500	20.0	1.2	1.1	1.3	1.8	1.8
17,500-22,600	8.2	1.5	1.4	1.0	1.4	1.9
22,600-35,500	3.2	2.6	2.5	2.0	0.9	2.3
35,500-92,000	1.7	3.3	3.3	2.5	0.6	2.7
92,000 and over	0.6	4.5	4.5	3.4	0.5	3.3
Total	100.0	1.8	1.8	1.8	1.8	1.8

[a] Based on Musgrave et al., forthcoming.

[b] Income, estimated at $772 billion for 1968, is defined to include money income plus corporate profits before tax less dividends, plus imputed income, wage supplements, interest on insurance, and capital gains.

[c] Of total property tax revenue of $29.9 billion, an estimated $13.8 billion is derived from residential housing, including $9.6 billion from owner-occupied homes and $4.2 billion from rental property.

[d] Tax burden allocated in line with the distribution of all capital income.

[e] Same as Column II plus effective net excise rate. The total amount of the excise burden or subsidy is set at $10 billion and the net rate is obtained as the excess of the loss (computed by allocating $10 billion in line with housing expenditures) over the gain (computed by allocating a similar amount in line with other consumption expenditures). In allocating the loss component, $7 billion is assigned to owner-occupied housing and $3 billion to rental housing.

[f] The tax on owner-occupied housing is allocated in line with all capital income and the tax on rental housing is allocated by rental expenditures.

[g] Allocated in line with housing expenditures, with $9.6 billion distributed by owner-occupied housing and $4.2 billion by rental housing.

[h] Obtained by allocating $13.8 billion in line with distribution of federal income tax burden.

other income, or income versus consumption) but in whether an initial level of tax-free base is allowed for. Use of an effective device (be it a credit, vanishing exemption, or form of circuit breaker) is thus of major importance for the property tax under all these assumptions. At the same time, it may be noted that the appropriate design of such allowance will differ depending on which incidence hypothesis is chosen.

At the other end of the scale the property tax is clearly an effective way of reaching high incomes. Yet an *in rem* tax on real property is hardly a sensible way of reaching capital income in the higher brackets. A fuller taxation of such income under the individual income tax, or else a personal tax on net worth (combining housing with other parts of the balance sheet), would surely be a superior approach.

B. *Gross vs. Net Burden*

It must be noted, however, that the burden distribution of the property tax differs greatly depending on whether it is seen in gross or net terms. Since the property tax may be deducted from taxable income under the federal income tax, the net tax in most cases is less than the gross tax. And since income tax savings per dollar of property tax rise with the taxpayer's income, the ratio of net to gross property tax declines as the taxpayer moves into higher rate brackets. To restate Column II of Table 1 in net terms would require no change at the bottom of the scale, since such households are not subject to individual income tax, but the net burden at the upper end of the scale might drop to as little as 1.3 percent. Since it is the net burden that matters if we take a national view of the tax structure, patterns such as those given in Table 1 (and generally used in burden estimates) exaggerate the progressive nature of the burden imposed by such taxes as the property tax, just as they understate the regressive nature of other taxes such as the sales tax. As far as *net* burden goes, the only effective way to strengthen the progressivity of state and local taxes is to disallow their deduction or to substitute a credit for a deduction approach under the federal income tax.

C. *Annual vs. Lifetime Data*

The above results were based on budget patterns drawn from annual cross-section data. The question remains how the patterns would differ if lifetime data were used instead. There is reason to expect that the use of lifetime patterns would render the burden distribution less regressive or more progressive. As regards allocation by capital income, it seems reasonable to expect that the high ratio of capital income among low income recipients reflects the role of the aged, so that the importance of capital income for low-income families would be less if viewed on a lifetime basis. Regarding allocation by housing expenditure, it also appears that the income elasticity of demand for housing is higher on a lifetime than on an annual cross-section basis (Frank de Leeuw and S. J. Maisel and A. Burnham). Once more use of lifetime data would render the distribution more progressive or less regressive.

Even if it turns out that allocation by lifetime data (under either the Column II or Column V interpretation) renders the tax less regressive for the lower half of the population, it is not obvious which should be considered the relevant finding. On the one hand, there is an obvious case in tax equity for relating progressive rates to a longer period than annual income, with lifetime averaging the ultimate extension of this principle. On the other hand, the lifetime view can be carried too far. While the aged poor should not be given preference over other poor who are not aged, I would hesitate to argue that, since the aged poor should have known better in distributing the consumption pattern of their lifetime earnings, cross-section regressivity poses no problem.

However this may be, the property tax will be with us for a long time and will continue to provide a major share of state and local tax revenue. The issue, therefore, is one of reform rather than replacement, and in this context the design of an effective policy of low-income relief is the major problem. Given progress in this direction, especially if combined with substitution of a federal credit for deduction, I would concur that the property tax on housing should be transferred from the regressive to the progressive column.

REFERENCES

M. S. Feldstein, "Tax Incidence in a Growing Economy with Variable Factor Supply," *Quart. J. Econ.*, forthcoming.

F. de Leeuw, "The Demand for Housing: A Review of Cross-Section Evidence," *Rev. Econ. Statist.*, Feb. 1971, *53*, 1–10.

A. Harberger, "The Incidence of the Corporation Income Tax," *J. Polit. Econ.*, June 1962.

M. Krzyzaniak, "The Long-Run Burden of a General Tax on Profits in a Neo-Classical World," *Pub. Fin.*, 1967, *22*.

S. J. Maisel and A. Burnham, "The Demand for Housing: A Comment," *Rev. Econ. Statist.*, June 1971, *53*, 410–13.

P. Mieszkowski, "The Property Tax: An Excise or a Property Tax?" *J. Pub. Econ.*, Apr. 1972.

R. A. Musgrave, J. J. Carroll, L. D. Cook, and L. Frane, "Distribution of Tax Payments by Income Groups, a Case Study for 1948," *Nat. Tax J.*, Mar. 1951.

———, K. E. Case, and H. Leonard, "The Distribution of Fiscal Burdens and Benefits", *Pub. Fin. Quart.*, forthcoming.

K. Sato, "Taxation and Neo-Classical Growth," *Pub. Fin.*, 1967, *22*.

DISCUSSION

HARVEY E. BRAZER, University of Michigan: Henry Aaron finds in the Harberger-Mieszkowski approach a better theoretical foundation than ever before for estimates of property tax incidence. The basis of Mieszkowski's theoretical foundation consists of a fixed capital supply assumption and the assumption that industries are specific to particular taxing jurisdictions, while at the same time capital, as well as labor, is perfectly mobile. Then, if one begins by assuming that the property tax is a uniform national tax on all real capital (including land), the conclusion that it will be borne by owners of capital through a reduction in net returns on capital follows. (It seems to me to be at least as plausible to assume that labor is in fixed supply while capital supply is yield-elastic, and conclude that the property tax reduces real wages.)

If one abandons the uniformity assumption and instead allows for local variations in the tax rate, above and below the "national average," the tax will be found to produce "excise tax" effects. The high tax communities experience higher prices while the lower-than-average tax rate communities enjoy lower than average prices. Thus, some households gain *qua* consumers while others lose, due to variations in local tax rates, but the property tax reduces net returns on capital by the amount levied. Thus the tax is as progressive as the distribution of the ownership of capital, modified by the income distribution of excise tax effects.

Surely this analysis is just a bit troublesome. I do not know of any good reasons why many industries should be specific for very long with respect to particular taxing jurisdictions. And with capital perfectly (highly?) mobile, I should not expect either the costs of capital or the prices of products to be higher in high-tax than in low-tax jurisdictions (other things being equal). Even more elusive, to my mind at any rate, is the significance of the average rate of tax, once the specificity of industries to given jurisdictions is dropped. The driving force, if there is one, must be the difference in tax rates between relevantly alternative locations, rather than the local rate relative to the average rate.

A satisfactory theory of the incidence of the local property tax in its peculiarly American institutional setting must recognize that the very essence of this tax is that it is not one tax but many taxes, each with a different effective rate. Thus I have considerable difficulty getting past Richard Musgrave's introduction. It is perfectly all right for Musgrave to address himself to "the incidence of a hypothetical national tax." But it is not enough merely to assert that local jurisdictional differentials "are not central to the issue of progression versus regression." Quite the contrary, I can imagine the development of an analysis in which a uniform national tax turned out to be regressive while, within the same analytical framework, a local tax with wide rate differentials proved to be progressive. At any rate, I believe that more needs to be made of the consequences of the availability, for capital and labor, of lower-tax alternative locations.

With respect to the property tax levied on industrial property, Musgrave opts for partial forward shifting through the medium of cost-plus administered prices. But the taxed firms will in many, if not most, instances be national (or even multinational) in character, producing any one product in plants in several localities, each with a different effective property tax rate. As a consequence, some difficult questions are raised for this route to analysis of incidence. Which property tax rate does the firm take into account in setting its administered price? And what of the tax rates paid by those firms that are price takers rather than price setters? The case for forward shifting seems to me to be extremely weak. Clearly it can be only that tax rate that the price setter adds to his costs that matters for forward shifting. The property tax paid by all other firms in the industry cannot affect prices and presumably not wages, so it becomes a profits tax, as does any property tax paid by the price setter other than that which governs his pricing decision.

Thus, I am more or less inclined to the belief that tables such as Aaron's Table 2 and Musgrave's Table 1 might best be avoided. With considerable hesitation I am tempted, rather, to prefer tables showing the distribution among income groups of tax payments (or imputed payments of renters) on residential property. For business and industrial property, with much timidity, I would accept the assumption that property taxes come out of profits *in the short run*, and distribute these according to capital income. Such tables may not present the true incidence picture, but there is some appeal, I find, in holding to the view that a tax stays put unless the contrary can be proven.

DICK NETZER, New York University: I do not think it is possible to make reliable empirical statements about the burden distribution of "the" property tax on a nationwide scale. One can view the property tax as a collection of partial capital taxes generating the excise tax effects (including *both* the backward and the forward shifting that is possible with excise taxes) expected of partial capital taxes or, alternatively, as a collection of partial commodity taxes with a fair amount of the burden borne by owners of land and capital. I think that one cannot usefully view the property tax as a general tax on capital with incidental variations around the burden distribution to be expected from a uniformly applied general tax, with the argument confined—as in these papers—to the extent of market imperfections.

Both Henry Aaron and Richard Musgrave do just that. The findings of an analysis based on such premises—that the property tax is not generally regressive in burden distribution (Aaron), that the tax is not an effective way of reaching higher income households (Musgrave), or that a personal exemption could produce a much more satisfactory burden distribution (Musgrave)—are just the beginning, not the end, of the work necessary to make policy-relevant statements about the equity and efficiency characteristics of this fiscal instrument.

I say this, first, because fundamental policy decisions are made for fifty separate state systems, and tax rate determinations are made by more than 70,000 noncoterminous local government units and "taxing areas." Second, and perhaps more relevant to this discussion, the property tax *is* the only tax for which there exist more than 70,000 different legal tax rates, hundreds of thousands of distinctive effective rates (because of assessment nonuniformity) and probably (when all the overlying local taxing units are added up) several million distinctive effective tax rate combinations. Knowledge of mobility and elasticities is of the essence in any appraisal of burden distribution in particular places, to the extent that the tax is *not* a uniform, general tax. Thus, I argue that, to be at all useful, empirical work on this subject must be specific to place and time and that the results of such work will exhibit wide departures from the global conclusions offered by both Aaron and Musgrave.

In seeking evidence to support their global conclusions, both authors stack the deck by making unduly sweeping (or just plain erroneous) statements about mainly empirical questions that are of some importance. For example, Aaron holds that gross rent multipliers increase systematically with rent; the only solid evidence which supports this shows that the multipliers are low indeed for slum property, but there is no support for the proposition in regard to the rest of urban housing markets. There is a similar problem with regard to the quality of assessment administration. We know that the worst housing in some (but not all) large central cities tends to be substantially overassessed. But Census evidence suggests that in the aggregate assessment discrimination is not massive or systematic and one can cite a fair number of cases in which widespread overassessment of high-value property (e.g., newer single-family and apartment houses) is the rule. Again, I rather suspect that up-to-date evidence in regard to consumer expenditure for "low-tax consumption" items (Aaron, Table 3) would show a rather different pattern—a more progressive one—for these forms of consumer expenditure. Aaron and Musgrave err, in my view, in opposite directions with regard to the measurement of income, Aaron by uncritically accepting that permanent income is the proper measure and Musgrave by accepting data on the bottom brackets of the income distribution that are, to say the least, suspect. I believe Musgrave is quite wrong in his appraisals of the supply and demand elasticities for low-quality housing; the evidence suggests the opposite conclusion. He is also incorrect, as the earlier comment suggests, in stating that low-income rental housing tends to be underassessed.

Although I would reject Aaron's rule of thumb, that the property tax should be considered progressive unless proven otherwise, I am on balance more sympathetic to his concluding statement that a shift away from the property tax "should not be justified by the view that its burdens are distributed regressively" than I am to Musgrave's argument that the property tax is not "a sensible way of reaching capital income in the higher brackets." Of course it is not, but few argue otherwise. The proper question is, instead, whether this existing tax is reasonably acceptable on equity and efficiency grounds—in specific places and with marginal reforms that (a) are not themselves objectionable and (b) have some chance of acceptance.

Finally, it is worth noting briefly some evidence on the pervasiveness of excise tax effects of the type that arise from nonuniformity (rather than imperfect competition). Using data for 1966–67, I have estimated that the mean effective rate on all capital subject to the tax was 1.73 percent; the rate for total tangible national wealth was 1.02 percent; and the rate for national wealth exclusive of publicly and institutionally owned capital was 1.34 percent. This, of course, reflects the large portion of national wealth that escapes the property tax entirely. Thus, a very large part of taxed capital must be taxed at above-average rates, with excise tax effects the rule. We have relatively little evidence on the dispersion of effective property tax rates on individual parcels, except for housing. Housing data suggest that 80 percent or more of all housing units in urban areas are subject to effective rates above the average for all privately owned capital. This is also probably true of nearly all commercial structures, a large proportion of industrial structures in larger urban places, most public utility property (where this is subject to the property tax at all) and most mining property. This preponderance of positive excise tax effects is the empirical basis for my conclusion that our authors start off on the wrong foot in seeking generalizations about the global incidence of the property tax.

ANN F. FRIEDLAENDER, Boston College: I am uneasy about accepting the Harberger-Mieszkowski analysis as the proper vehicle to analyze the incidence of the property tax or, for that matter, any tax on capital. My reservations have to do with the nature of factor supplies, the government's equilibrating adjustment, and the initial tax structure.

There are two primary ways in which static analyses of capital taxation may overstate the burden borne by capital. First, as Martin Feldstein and others have indicated, if savings fall as a result of capital taxation (whether through income, interest rate, or wealth effects), capital may bear a much smaller proportion of the burden in the long run than in the short run. Thus insofar as the property tax is a tax on capital, the owners of capital may actually bear much less of the tax than Henry Aaron or Richard Musgrave suggest. Second, if capital is variable, it becomes necessary to distinguish between the land and improvements component of the property tax. If land is fixed and capital is not, it is not difficult to tell a plausible story in which land bears the full burden of the property tax in the long run. Hence treating the property tax as a tax on capital alone may be very misleading in terms of long-run incidence. While it is certainly true that the long run may be very long indeed, it is also true that property and capital taxation involve dynamic incidence effects that should not be ignored.

Second, I would like to stress the role of the government's equilibrating adjustment. Most incidence analyses, including Harberger-Mieszkowski's and Musgrave's paper presented here, implicitly or explicitly assume that any change in a given tax is accompanied by a change in a neutral head tax sufficient to keep the government's budget in equilibrium. In a world of fixed factors, many kinds of taxes fulfill the neutrality condition, and whether head taxes are in fact available is not terribly important. But in a world of variable factor supplies, neutral taxation is probably impossible to achieve, and the incidence effects of a change in sectoral capital taxation will be quite sensitive to whether the equilibrating adjustment is, say, a change in a wage tax or a change in general commodity taxation. Thus, any empirical analysis of the property tax should explore the equilibrating adjustments actually open to the government in more detail.

Third, I think that the role of the existing tax structure has not received sufficient attention in incidence analyses. Harberger and Mieszkowski assume that there are no initial distorting taxes. But it can be shown that, in the absence of initial distorting taxes, a distorting change in taxation creates no first-order dead weight loss. Even though relative factor and commodity prices, consumption and production may change in response to the tax change, consumers will stay at their initial level of indifference. Thus, in a very real sense, there are no incidence effects associated with sectoral taxation in the Harberger-Mieszkowski world.

For a dead weight loss to exist, there must be initial distorting taxation, as most certainly exists in the United States. But in the presence of initial distorting taxation, the change in welfare resulting from a change in sectoral taxation is often indeterminate, as is the direction of the relative change in factor prices. To determine the full static incidence of capital taxation, we not only need data on elasticities of supply, demand, and factor substitution, but also data on the initial tax structure and income effects. Thus, even in the short run, capital may bear a much smaller burden of capital taxation than that indicated by Harberger and Mieszkowski.

I conclude that both static and dynamic analysis indicate that there are good reasons to question the presumption that a sectoral tax on capital is fully borne by capital. Consequently, until we have firmer theoretical foundations upon which to measure the short-run and long-run incidence of the property tax, I will remain agnostic about the question of its progressivity or its regressivity.

EARL R. ROLPH, University of California, Berkeley: Henry Aaron's prospective research strategy on the incidence of the property tax appears to have three features. The property tax is viewed as having an average rate effect, found by comparing total property tax liabilities with the universe of the value of all real assets or of all privately owned real assets. From this approach, it follows that the tax rates will display dispersion effects and this aspect in turn is expected to give rise to excise tax effects. I shall restrict my comments to the second feature—the dispersion effects.

If property is taxed more heavily in community A than in community B, does it follow that property will tend to move to community B. as Aaron claims? If only the tax rate matters and the cost of movement is zero, Aaron's expectation is appropriate. But if the higher taxes in community A are accompanied by a more attractive menu of government services, just the opposite may result. Aaron's approach would make sense if the American property tax were a national tax. But the property tax is a local tax and as such it is closely related to local government programs and to benefit features generated by such programs. Even in cases where the benefit element may appear to be small, as in the case of some types of industrial facilities, the property tax may, indirectly, be a benefit levy, compensating local residents for the negative externalities commonly associated with such facilities. To proceed as if these features can be safely ignored seems to me to make a serious mistake in research strategy.

Turning now to Richard Musgrave's paper, I detect a greater reluctance on his part to assume that the property tax is shifted to consumers, and for this stance, let me be the first to congratulate him. He retains, however, the peculiar Musgravian theory of how markets really work or, should I say, do not work. We are informed, for example, that "Housing markets especially for low-income housing in central city settings may well be markets where landlords act as restrained monopolists . . . who find it justified to raise rents as taxes are increased. . . . "

When landlords number in the thousands, this theory of pricing is difficult to swallow. Musgrave wants us to believe that "restrained monopolists" would not ask for higher rents when demand-supply considerations would so indicate as long as property taxes do not increase.

Musgrave's paper also assumes that the property tax is to be examined as a hypothetical national tax. As already observed, I believe such an exercise to have merit but at most only as the beginning of the story. Viewed as a local tax, the incidence of the property tax is ambiguous. How does one distribute a tax liability on a warehouse located in an industrial enclave such as Emeryville, California? Or, in the case of residential enclave, how should we interpret the property tax liability levied on a dwelling owned by a university professor? Statements that property tax liabilities are "regressive" or "progressive" presuppose that it is permissible to neglect benefit features that affect rental values of properties. It seems to me that research efforts might be directed to attempting to estimate the "net" tax. If it turns out this is a positive number, we would at least know that there is something to distribute.

George E. Peterson, The Urban Institute: The advances in incidence theory that Henry Aaron and Richard Musgrave summarize show that the property tax is a less regressive instrument than was commonly supposed a short time ago. Aaron and Musgrave suggest that geographical rate differentials likewise may not greatly disturb the general conclusions about the tax's regressivity that are drawn from looking at the property tax as a uniform nationwide tax. It is this aspect of the argument on which I want to focus.

First, there is a factual question: how do local property tax rates vary with household income levels? Aaron's unpublished regressions seem to show that for the most part tax rates and household income increase *pari-passu*. However, the data used in these regressions refer to jurisdictions. His results show only that communities with low median incomes by and large pay lower property tax rates than communities with high median incomes. The proper comparison for an incidence study would give the average tax rate faced by *households* at different income levels. Such household stratification would reveal a markedly more *regressive* structure of local tax rates than do the jurisdictional regressions he reports. Thus, Aaron's conclusion that geographic rate differentials on balance increase the progressivity of the property tax's differential incidence does not seem to be warranted.

Of more interest in judging the present property tax *system*, however, is the tax's balanced-budget incidence, i.e., the distribution of expenditure-controlled property tax rates, rather than actual tax rates. It is the local *capacity* to raise property taxes that is relevant to a judgment on whether the system of local property taxation is more or less regressive than a uniform nationwide property tax would be. Data which I have collected for several states show that, as the average level of local property taxation increases, the incentive to adopt fiscal zoning increases, and the stratification of communities by household income and tax base becomes more pronounced.

Under what conditions will the higher (expenditure-controlled) tax rates associated with low-income communities be passed on to residents there? As Aaron notes, if both labor and capital are fully mobile among the various jurisdictions in a metropolitan area, tax differentials will be fully incorporated into property values—i.e., balanced-budget rate differentials will not be passed on to the consumers of housing. Suppose, however, that we view low-income and high-income housing as two distinct industries. Data for many metropolitan areas indicate that low-income housing (especially new additions to the low-income housing stock) are highly concentrated in low-tax base, high (expenditure-controlled) tax rate areas. Moreover, existing zoning regulations and other land use constraints frequently make it impossible for a significant volume of low-income housing to be constructed *except* in low-tax base jurisdictions. In other words, capital for the low-income housing industry in many instances emphatically is *not* mobile among communities, due to the severe land use restrictions that have been imposed. The effect is to force low-income housing as an industry to produce its services under a higher effective tax rate structure than high-income housing. If local labor supplies are fully mobile, then the higher average tax conditions that burden the production of low-income housing will be passed on in their entirety to the consumers of this product. I am not in a position to hazard a guess as to the quantitative importance of the forward shifting of property tax burdens to consumers of low-income housing that occurs in response to fiscal zoning, but I would imagine it to be large.

Finally, when judging the balanced-budget incidence of the *local* property tax system, it is imperative to keep in mind the effect of dynamic changes. To my mind, the most cogent case that can be constructed on regressivity grounds against the system of local property taxation is based on the steady deterioration that is occurring in the tax bases of low-income jurisdictions *relative* to the tax bases of high-income jurisdictions. In all of the older standard metropolitan statistical areas, where reliance on property taxation is greatest, low-income jurisdictions have suffered a marked deterioration since 1960 in the size of the local tax base, as compared to wealthy jurisdictions. Even if the steadily worsening condition of low-income communities' relative position is capitalized into housing values, I am convinced that this change has a continuing regressive effect, inasmuch as the capital losses to a considerable extent are borne by the low-income owners of property there.

Like Aaron and Musgrave, I believe that the property tax is much less regressive than it has been portrayed. Viewed from the perspective of balanced-budget incidence, however, local tax rate differentials certainly make the *local* property tax system more regressive than a uniform nationwide property tax would be. And I cannot agree that a study of differential tax incidence, cast in a national framework, provides nearly as useful a way of looking at the effects of local rate differentials as do balanced-budget comparisons.

WOMEN AT WORK: AT HOME AND IN THE MARKET

Recent Trends and Long-Run Prospects for Female Earnings

By VICTOR R. FUCHS*

This paper discusses the recent behavior of the female-male earnings differential and asks how it is likely to change in the future.[1] After adjusting for hours, age and schooling, we find that, contrary to results obtained in other studies from unadjusted data, the female-male earnings ratio among whites *increased* between 1959 and 1969 by 4.8 percent despite an unprecedented increase in the female-male *employment* ratio among whites of almost 20 percent in the same decade—a favorable augury for future trends in the earnings ratio.

The data used below come from the 1/1000 sample of the 1960 and 1970 Censuses of Population.[2] Due to lack of time and space, I focus on the sex differential among whites. However, the color differential between blacks and whites narrowed appreciably between 1959 and 1969. Black female hourly earnings, adjusted for age and schooling, rose 82 percent compared with 68 percent for black males and 53 percent for white females. By 1969, hourly earnings of black females were only 15 percent less than those of white females of comparable age and schooling, while for women with more than twelve years of schooling the adjusted color differential had practically disappeared.

Total earnings in 1959 (1969) are calculated for workers classified for a variety of characteristics; total annual hours are estimated by multiplying weeks worked in 1959 (1969) by hours worked in the census week in 1960 (1970) for each worker and summing across all the workers in a classification. Finally, total earnings are divided by total hours to obtain, for each classification, average hourly earnings—or, equivalently, the mean of individual average hourly earnings weighted by annual hours. I focus on hourly earnings as the best mea-

---

* Professor of economics, City University of New York, and Vice-President, Research, National Bureau of Economic Research (*NBER*). Work on this paper was begun while I was a Fellow at the Center for Advanced Study in the Behavioral Sciences, 1972–73, with partial support provided by the Russell Sage Foundation. It was continued at the *NBER* under a grant from the Rockefeller Foundation. Carol Breckner, Phyllis Goldberg, Jan Platt, and Christy Wilson provided research assistance and many useful suggestions; Charlotte Boschan wrote the earnings program; Robert Michael and Yoram Weiss made helpful comments on an earlier draft. I am grateful to all of the above; they bear no responsibility for the views expressed. In particular, this paper has not been submitted to the *NBER* Board of Directors for approval and is not an official *NBER* publication.

[1] I do not consider determinants of the general level of earnings (female and male).

[2] See Fuchs (1968) for a discussion of the strengths and shortcomings of this source, which provides much useful information on employed persons' sex, schooling, age, race, marital status, occupation, and industry. All persons who were at work during the census week and had earnings in the year preceding the census are included in the sample except agricultural, religious, and unpaid family workers, since, as is well known, the estimation of earnings and work hours for such workers is difficult.

sure of the price of labor from the standpoints of both supply and demand.[3]

What are the prospects for female earnings? In order to answer this question one needs a clear understanding of the factors that account for the sex differential as well as the ability to predict how these factors will change over time. In my view, neither inherent physical or mental differences nor employer discrimination can explain most of the differential. (See Fuchs, 1971, Jacob Mincer and Solomon Polachek, and Burton G. Malkiel and Judith A. Malkiel.) This is not to deny the existence of some discrimination by many employers and a great deal by some, but those who discriminate do pay a price (Gary S. Becker). For most employers the desire for profit or the fear of loss makes them unwilling or unable to absorb the 30 to 40 percent difference in labor costs that is implied if the differential in earnings is attributable only or principally to employer discrimination.

The major explanation, it seems to me, is *role differentiation*, which begins in childhood and eventually affects labor force attachment, choice of occupation, location and hours of work, postschool investment, and consumer and fellow employee attitudes. This role differentiation was functional at a time when men worked long hours at heavy jobs in mining, manufacturing, transportation, and construction while women specialized in work at home including the bearing and raising of many children. Such differentiation is less functional now, and much of the recent tension regarding sex roles probably arises from the lagged adjustment of the law, customs, and institutions to technologic and economic changes. These changes include sharp reductions in infant and child mortality, dramatic improvements in birth control, and major shifts in requirements of the job market.

The effects of these changes on female earnings for given labor market productivity can be analyzed with the familiar tools of demand and supply, supplemented by attention to changes in the relative market productivity of men and women that are not captured by adjustment for age and schooling. In this analysis price is the wage of females relative to males, and quantity is employment of females relative to males.

Because the sexes are not perfect substitutes, the demand curve is not completely elastic; it is probably becoming more elastic over time. The demand curve shifts over time as a result of changes in the industrial and occupational mix of the economy. In particular, the growth of a service economy and the decline in the importance of heavy manual jobs tend to move the demand curve to the right. Demand also is affected by the removal of legal and institutional barriers to women and by the greater acceptance of women in a variety of occupations and roles by consumers and other employees.

The relative supply curve is dominated by changes in female labor force participation because the supply of male labor tends to be fixed. Shifts in the curve are related to the decreases in infant and child mortality, the improvements in birth control, the increase in the absolute level of wages, and the desire of women to achieve greater autonomy through maintenance and enhancement of labor market skills.

During the 1960's there was a very large increase in female employment, which I interpret as primarily a shift in the relative supply curve. This shift tends to depress average female earnings in the short run, not only because increased quantity lowers price but also because the new entrants tend to have less schooling and less labor market experience. As Reuben Gronau has shown, they are probably less able (in labor market terms) than those already at work.

---

[3] However, the analysis can easily be extended to a consideration of annual earnings, since on average employed women worked about three-fourths as many hours per year in 1970 as employed men.

In this same decade, however, there was increased demand for female labor due to the rapid growth of industries such as health and education that traditionally have been large employers of women. The beginning of legal and institutional changes within industries and occupations also contributed to the increase in demand.

The net effects of these shifts on female earnings are presented in Table 1. The female-male earnings ratio adjusted for age and schooling (R) is calculated as

$$R = (F/\hat{F} + \hat{M}/M)/2$$

where

F = average hourly earnings of females,
M = average hourly earnings of males,
H = total annual hours worked by females,
K = total annual hours worked by males,
subscripts a and s = age group a and schooling group s,

TABLE 1—SEX DIFFERENTIALS IN HOURLY EARNINGS, WHITE NON-FARM EMPLOYED, 1959 AND 1969

	Average Hourly Earnings, 1969		Female-Male Ratio, Hourly Earnings Adjusted for Age and Schooling		Percent Change in Female-Male Ratio of Adjusted Hourly Earnings
	Females	Males	1959	1969	1959 to 1969
All	2.70	4.46	.61	.64	4.8
Northeast	2.92	4.71	.63	.66	5.2
North Central	2.68	4.52	.61	.62	2.2
South	2.41	3.98	.60	.63	5.8
West	2.89	4.78	.60	.64	6.4
≤12 Years of Schooling	2.41	3.84	.61	.62	1.5
>12 Years of Schooling	3.44	5.75	.59	.66	11.4
<35 Years of Age	2.64	4.19	.71	.74	3.1
≥35 Years of Age	2.79	4.84	.57	.59	4.2
Married, Spouse Present	2.69	4.67	.59	.61	3.0
Never Married	2.72	3.06	.81	.86	6.0
Other	2.71	4.04	.65	.69	6.0
<35					
Married, Spouse Present	2.64	4.46	.73	.70	−3.4
Never Married	2.59	2.90	.81	.86	6.1
Other	2.73	3.87	.71	.80	13.0
≥35					
Married, Spouse Present	2.76	4.94	.56	.57	3.0
Never Married	3.25	3.86	.78	.80	2.9
Other	2.69	4.22	.64	.67	4.5
Private Wage and Salary	2.52	4.27	.59	.62	5.2
Government	3.43	4.44	.79	.77	−2.4
Self-employed	2.71	5.68	.51	.57	11.4
<35					
Private Wage and Salary	2.48	4.02	.71	.73	2.2
Government	3.29	4.26	.82	.84	3.2
Self-employed	2.74	5.51	.73	.69	−5.4
≥35					
Private Wage and Salary	2.59	4.68	.54	.57	5.1
Government	3.64	4.66	.76	.72	−5.4
Self-employed	2.69	5.81	.50	.57	13.9

Source: 1/1000 Sample, 1960 and 1970 *Census of Population*. Calculations by author.

$$\hat{F}=\sum_{as}(M_{as}H_{as})/\sum_{as}H_{as} \text{ and}$$
$$\hat{M}=\sum_{as}(F_{as}K_{as})/\sum_{as}K_{as}.$$

It is, therefore, an average of the results obtained by standardizing female hours on male wage rates and male hours on female wage rates across forty-nine age-schooling cells. The percentage change in the ratio from 1959 to 1969 is $(100)(R_{70}-R_{60})/R_{60}$ (calculated from unrounded data).

The table reveals that the adjusted female-male earnings ratio for all whites *increased* between 1959 and 1969 from .61 to .64, a gain of 4.8 percent. As suggested above, this increase, although small, was noteworthy given the unprecedented increase in the female-male *employment* ratio during the same decade. The improvement in female earnings was evident in all four regions, with all except the North Central showing gains of over 5 percent.

When the sample is disaggregated by schooling, we find that the rise in female earnings was very large for those with at least some college; the sex earnings ratio jumped 11.4 percent. For workers with twelve years of schooling or less, the increase was only 1.5 percent. The explanation for this difference lies, I believe, in an exceptionally rapid growth of *demand* for well-educated women and a relative increase in *supply* of less-educated women.

The demand phenomenon is illustrated in Table 2, which lists major industries and occupations in order of their rates of growth from 1960 to 1970. It is evident that those industries and occupations which were large employers of well-educated women in 1960 were precisely the ones that experienced the largest increases in demand for labor between 1960 and 1970. A simple projection based on 1960–70 industry group growth rates and 1960 sex-schooling distributions by industry group reveals an expected increase in demand between 1960 and 1970 of 47 percent for well-educated females compared with 28 percent for well-educated males and 20 percent for less-educated females.

On the supply side two aspects must be considered: changes in the number of females in the population relative to that of men at the two schooling levels and changes in the relative number of females employed. We see in Table 3 that although the female-male ratio rose more rapidly for the less educated in both cases, the differential between schooling groups was particularly striking for employment. Why did the relative employment of less-educated women grow faster than that of the more educated when the latter's relative wages were increasing more rapidly?

It is probably true that the relative supply curve of the less educated is somewhat more elastic than that of the well educated. The principal explanation, however, is differential shifts in the female labor supply in response to increases in the *absolute* level of earnings (and other reasons) rather than movements along relative supply functions in response to changes in *relative* earnings. An increase in the general level of wages has very little effect on male labor force participation rates at any level of schooling. It has some effect on well-educated females but a greater relative effect on the less-educated ones because the well-educated females are already at a higher level of participation. Thus the female-male relative supply function shifts more for the less educated in response to a rise in the general wage level.

Disaggregation by age reveals that the increase in female earnings was about the same for those below thirty-five as those above, but further disaggregation by marital status or class of worker reveals that this is the result of conflicting trends. When females and males are compared by marital status, the smallest decrease in the sex earnings differential is found for married persons. This is not surprising given the huge increase in the labor supply of

TABLE 2—EMPLOYMENT CHANGE 1960 TO 1970 AND COLLEGE-EDUCATED FEMALES AS PERCENT OF LABOR FORCE, 1960, BY MAJOR INDUSTRY AND OCCUPATION

	Employed (thousands)		Employment in 1970 as Percent of 1960	College-Educated Females as Percent of Labor Force in 1960[a]
	1960	1970		
Major Industry Group				
Professional and related services	7,695	12,780	166	29.8
Business and repair services	1,607	2,253	140	5.0
Finance, insurance, real estate	2,695	3,652	136	8.4
Public administration	3,086	4,056	131	7.9
Wholesale and retail trade	11,793	14,613	124	3.8
Entertainment and recreation	503	591	117	5.6
Manufacturing: durable	9,833	11,124	113	1.6
Construction	3,816	4,219	111	0.7
Transportation, communication, public utilities	4,458	4,906	110	2.5
Manufacturing: nondurable	7,681	7,756	101	2.4
Mining	654	605	93	1.4
Personal services	3,862	3,294	85	3.5
Major Occupation Group				
Professional and technical	6,986	10,831	155	27.0
Clerical and kindred	9,126	13,035	143	12.9
Service except private household	5,754	8,065	140	2.8
Sales workers	4,637	5,267	114	4.1
Craftsmen, foremen	8,945	9,996	112	0.3
Operatives	11,347	12,582	111	0.7
Managers and administrators	5,626	6,139	109	3.4
Laborers except farm	3,322	3,213	97	0.1
Private household workers	1,718	1,093	64	—

[a] For industries, experienced civilian labor force over fourteen years of age; for occupations, over twenty-five years of age.

Sources: 1970 Census of Population, General Social and Economic Characteristics, PC(1)-C1 U.S. Summary, Tables 81 and 82; and *1960 Census of Population*, Industrial Characteristics, PC(2)-7F Table 21 and Educational Attainment PC(2)-5B Table 8.

married women during the decade.[4] The female-male labor force ratio increased 32 percent from 1960 to 1970 for married persons. The increases in the ratio for never married and "other" were 6 percent and 10 percent respectively. The increase in labor force participation of married women was particularly great among those under age thirty-five, with a noticeable effect on the sex earnings differential as shown in Table 1. Of the six marital status-age classes, only married persons under thirty-five showed a relative decrease in female earnings. Indeed, for the never married and "other" the increase was greater for those under thirty-five than for those above that age.

When the comparisons are made by class of worker, we find that in government, where female earnings have been relatively highest, there was actually a small decrease in the earnings ratio between 1959 and 1969. Further disaggregation by age reveals that this decrease was concentrated in the over-thirty-five category. One possible explanation is that this is the result of the strong relation between earnings and seniority in government employment. Older women who entered government during the decade had less seniority than men of comparable age and therefore lower earnings. At younger ages seniority in government would be more comparable between the sexes, and in the private sector earnings are not so rigidly determined by seniority at any age.

---

[4] This is not because the labor market is segmented by marital status, but because the new entrants have less earning power than those already in the labor market.

One variable that highlights the effects of role differentiation on female earning power is employment status five years prior to the census. Of the white males twenty-five and over who were employed in 1970 and in 1969, only 6 percent were not at work in 1965. The comparable figure for white females was 24 percent. Among married white females age twenty-five to forty-four, more than 36 percent were not at work five years earlier. The comparable figure for males is 8 percent. These differences are noteworthy because there is a very large differential in hourly earnings, about 18 percent, between employed persons who were at work five years earlier and those of comparable sex, color, age, and schooling who were not. For white females in government this differential is almost 24 percent!

TABLE 3—FEMALE-MALE RATIOS OF POPULATION AND EMPLOYMENT BY LEVEL OF SCHOOLING, NON-FARM WHITES, 1960 AND 1970

	Female-Male Ratio		Percent Change in Female-Male Ratio
	1960	1970	1960 to 1970
Population 25–64			
$\leq$12 years	1.12	1.17	4
$>$12 years	.81	.79	−2
Employment 25–64			
$\leq$12 years	.47	.58	23
$>$12 years	.41	.44	7

Sources: for population 1960: U.S. Dept. of Commerce, Bur. of the Census, *U.S. Census of Population: 1960, I, Characteristics of the Population*, Part 1, U.S. Summary, Washington, Table 173.

For population 1970: U.S. Dept. of Commerce, Bur. of the Census, *U.S. Census of Population: 1970, Detailed Characteristics* U.S. Summary, Washington, Table 199.

For employment 1960: U.S. Dept. of Commerce, Bur. of the Census, *U.S. Census of Population: 1970, Subject Reports, Educational Attainment*, Washington, Tables 4 and 5.

For employment 1970: U.S. Dept. of Commerce, Bur. of the Census, *1970 Census of Population, Subject Reports, Earnings by Occupation and Education*, Washington, Tables 1 and 7.

This preliminary reading of part of the evidence on recent changes does not constitute a rigorous test of a theory of female wage determination, but I think one conclusion is warranted. If, during a period of rapid increase in supply, female earnings were more than able to hold their own and for some groups show significant gains, the long-run prospects for women must be viewed as favorable.

In the decades ahead female labor force participation is likely to continue approaching the male rate, and at some point the growth in the female-male employment ratio will taper off. Although in the short run the increased labor force participation of women tends to depress female earnings, in the long run it will raise them. Because they will expect to be in the labor force for a significant portion of their adult lives, women will be more career-minded while attending school, choosing an occupation, and investing in themselves after they leave school. Employer expectations concerning continuity will also change with important implications for job and training opportunities.

The women who entered the labor force in large numbers in the 1960's did not have much labor market experience by 1969. After the transition, however, the average work experience of employed women will increase. Moreover, when female labor force participation rates stabilize at a high level, new entrants will consist primarily of young women who will have been less exposed to role differentiation at home and in school than those now in the labor force.

The increasing acceptance of women in a variety of occupations, the narrowing of sex differences in experience and post-school investment, and a continued shift away from heavy manual jobs all augur well for female earnings.

REFERENCES

G. S. Becker, *The Economics of Discrimination,* 2d ed., Chicago 1971.

V. R. Fuchs, *The Service Economy,* New York 1968.

———, "Differences in Hourly Earnings Between Men and Women," *Monthly Labor Rev.,* May 1971, 9–15.

R. Gronau, "Wage Comparisons—A Selectivity Bias," Natl. Bur. of Econ. Res. working pap. 13, Oct. 1973.

B. G. Malkiel and J. A. Malkiel, "Male-Female Pay Differentials in Professional Employment," *Amer. Econ. Rev.,* Sept. 1973, *63*.

J. Mincer and S. Polachek, "Family Investment in Human Capital: Earnings of Women," *J. Polit. Econ.,* forthcoming.

Education and Home Production

By Arleen Leibowitz*

Valuable insights into the labor supply of married women have recently been gained by analyzing the problem in the household production context. Pioneered by Jacob Mincer, this approach sees women as choosing not simply between work and leisure, but between work in the home, work in the market, and leisure. While income affects the total amount of work, its division between home and market depends on wage rates, productivity in the home, and the price and availability of substitutes for the wife's labor in the home.

Participation of married women in the market has, of course, increased dramatically in the past thirty years; this increase in labor force participation has been accompanied by various changes in inputs and outputs of household work which have only recently come under detailed scrutiny. It appears that over a fifty-year period the total time devoted to work in the home by urban nonemployed women has remained virtually constant at fifty-three hours per week (Joann Vanek). But since employed women have always devoted less time to household production than nonemployed women, and since the proportion of women in the labor force has been rising, the average time devoted to household tasks by all women has been falling over the past fifty years.

One of the factors drawing women into the labor market has been their rising educational attainment. But what is the effect of the considerable rise in schooling levels on production within the home? First, it is important to understand how home production varies with schooling. Michael Grossman has recently found that a married man's health is positively related to his wife's schooling level. Lee Benham finds evidence to support his hypothesis that the earnings of married men are positively related to their wives' schooling. Robert T. Michael finds that education affects the efficiency with which contraception is carried on. In contrast to these papers, which relate a woman's education to various household production outputs, the present paper will try to determine how schooling affects one of the inputs of household production, namely, time. Time budget data will be used to determine how time allocation to various activities varies with schooling level.

I. Household Production and Education

A striking relationship found consistently by students of female labor force behavior (e.g., Glen G. Cain and William R. Bowen and T. Aldrich Finegan) is that more-educated women are more likely to be in the labor force. This relationship is even stronger when family income is held constant (since women with more education tend to have higher family incomes, which, *ceteris paribus*, reduces labor supply). The most widely accepted explanation for this association is that education has the nonneutral effect of raising the productivity of labor market time more than that of time spent in home production. Thus the "cost" of not being in the labor market rises, inducing women to work outside the home.

---

* Research associate, National Bureau of Economic Research, and visiting assistant professor, Brown University. This research has been supported by a grant from the National Institute of Child Health and Human Development to the National Bureau of Economic Research. The following paper is not an official National Bureau publication since the findings reported herein have not yet undergone the full critical review accorded the National Bureau's studies, including approval of the Board of Directors.

Moreover, better-educated women can also be expected to consume more leisure, if family income other than the wife's earnings is positively correlated with the wife's education. It follows that because women with more schooling spend a greater proportion of their lifetime in the labor market, they must normally spend a smaller proportion of their time in home production, as household production analysis would predict from their greater opportunity costs of time. As a result, other inputs (purchased goods and others' time) will be substituted for the wife's time in the production of most household outputs and, since the relative price of time-intensive outputs rises, consumption will shift to less time-intensive outputs.

Census data verify that the supply of labor to the market is usually greater the higher the level of schooling attained by the woman. But the census data also highlight an exception to this rule: between the ages of twenty-five and forty—that is, the ages when young children are in the home—more-educated women supply no more time to the labor market than other women (Leibowitz, 1972).

These labor supply profiles imply that if labor supply is determined as the result of a process which optimizes the utility of household production over the lifetime, child care must differ from other kinds of household outputs produced throughout the entire lifetime.

In what follows, I therefore do two things: first, I will show that time budget data (household production data) are consistent with the labor force data—that is, that both kinds of data show that more-educated women do devote less time to household production over the life cycle, but more time to child care. Second, I will try to explain why better-educated women allocate their time in this way.

II. Average Time Inputs to Household Activities by Education

Time inputs to various domestic outputs were calculated from time budgets of 1,296 families collected by Kathryn Walker of Cornell University.[1] Table 1 indicates that women with up to four years of high school spend as much or more of their own time as women who had attended at least one year of college in the two kinds of home production carried on throughout the life cycle—meal preparation and laundry work. This, and the fact that husbands of women with more schooling spend greater amounts of time in meal preparation, substituting their own time for their wives', is consistent with the greater price of time of more-educated women.

In spite of their greater price of time, however, more-educated mothers spend more time in child care, both overall and per child. This is particularly striking in educational care. This is not due to a substitution of the mothers' time for the fathers' or other persons', since husbands of more-educated women also spent more time with their children.

A study of time use in Indiana families in 1961–62 also shows that the greater the education of the wife, the more time she spent in child care (Leibowitz, 1973). Women with college degrees spent more than twice as many hours in child care as women with less than twelve years of schooling, 83 percent more time than high school graduates, and 59 percent more time than women with one to three years

---

[1] The data were collected for the research project on *Use of Time for Household Work* in the Department of Consumer Economics and Public Policy, New York College of Human Ecology, Cornell University. I am indebted to Walker and Irma Telling for providing these data.

TABLE 1—TIME INPUTS TO HOUSEHOLD ACTIVITIES

Activity	High Education Group	Low Education Group
Time Inputs to General Household Production (minutes over two days)		
Meal Preparation		
By Wife	153.41	154.39
By Husband	12.85	10.53
Laundry–Wife[a]	27.26	31.12
Total Time Inputs to Child Care		
Physical Care		
By Wife	129.51	116.40
By Husband	14.90	12.26
By Others	5.67	4.41
Educational Care		
By Wife	90.96[b]	79.17
By Husband	40.77[b]	31.51
By Others	44.74	36.38
Per Child Time Inputs to Child Care (minutes over two days)		
Physical Care		
By Wife	59.6	48.3
By Husband	6.9	5.1
By Others	2.6	1.8
Other Care		
By Wife	41.9	32.9
By Husband	18.8	13.1
By Others	20.6	15.1
Number of Children	2.17	2.41
Wife's Preference Rating		
Physical Care	6.78	6.84
Other Care	7.43	7.28

[a] Minutes over one day.

[b] Significantly greater at 1 percent level than low education group.

Notes: Sample sizes for meal preparation and laundry time inputs are 627 and 667 for high and low schooling groups. Since child-care averages are calculated only for families with children, sample sizes are 493 and 591, respectively.

"Physical care" includes time spent in bathing, feeding, and dressing children and in caring for a sick child. "Other care" was defined as "all activities related to the social and educational development of family members, such as: helping with lessons, reading to children, taking children to social and educational functions."

Preference ratings are averages of ratings of activities by homemakers on a scale from a low of 1 (extreme dislike) to 9 (like exceptionally).

of college completed. The greater time expenditures of more-educated women are seen within age categories of children as well as over all families.[2]

Frank Stafford and C. Russell Hill (1973a) report that the increase in home production time caused by the presence of a preschool child is twice as great in high socioeconomic status (*SES*) families as in low *SES* families. In addition, they find that education has a positive impact on mothers' time inputs to children. However, time inputs to other household tasks tend to fall with education (although there is a tendency for time inputs to rise slightly at the highest level). In the next section, the allocation of time within the household is examined to understand these differences by schooling level.

---

[2] I am grateful to Sarah L. Manning, Purdue University, Agricultural Experiment Station, Lafayette, Indiana, for providing me with these unpublished results.

III. Regression Analysis of Time Inputs

If the income elasticities of demand for home outputs all equaled unity, if the elasticities of substitution between time and goods in the production of home outputs were all equal, and if schooling affected the productivity of time at home and in the market equally, we would expect to find no relation between schooling and child-care time inputs. However, if schooling augments the productivity of time in the labor market relative to time at home—as suggested by the fact that better-educated women supply more time to the labor market—the relative price of time-intensive goods would be higher for better-schooled women.

It is widely believed that relative to other home activities care of preschool children is time-intensive. Thus better-educated women face a higher price of child care due to their higher opportunity cost of time. This is one rationale for the smaller number of children desired (and produced) by more-educated women (Robert Willis and Michael). Yet while the cost of child care is higher for better-educated women, they devote more time to child care, both in total and per child. As contrasted with other home output not concentrated in the child-rearing years, child care must thus be characterized by one or more of three factors:[3]

1. greater income elasticity;
2. smaller (larger) price elasticity and elasticity of substitution between time and goods relative to other household outputs, if education raises (lowers) market productivity relative to home productivity;
3. greater education-induced increases in home productivity relative to market productivity, given price and substitution elasticities.

To determine how these factors operate to counteract the relatively greater cost of child care for more-educated women, I undertook more detailed analysis of the Cornell data. Time inputs for each of four household activities were regressed on measures of the presence of substitutes (such as capital goods and time inputs by others), productivity factors (such as wife's education and age), and demand factors (such as family income and number of children in various age classes), which increase the marginal productivity of time in the home and should be positively related to time inputs.

In Table 2 we see that time inputs to meal preparation depend on demand factors and decrease monotonically with inincreasing age of children. The income proxies (rooms per capita and husband's education) are positively but not significantly related to time inputs. For each ten minutes devoted to meal preparation by the husband, the wife reduces her input to this task by five minutes—suggesting that husband's time in meal preparation is a substitute (albeit a highly imperfect one!) for the wife's time. Among the productivity variables, education has no significant impact on time inputs to meal preparation, but older women spend more time at this activity.

In laundry work, capital goods significantly affect the time required for a given number of loads washed or a given family size.[4] For example, the housewife who uses a nonautomatic washer spends thirty-nine minutes more in washing than one who uses an automatic washer. More-

---

[3] For the derivation of these results, see Leibowitz (1972).

[4] Regressions not shown.

TABLE 2—REGRESSION ANALYSIS OF TIME INPUTS TO HOUSEHOLD PRODUCTION

Time[a] Spent In:	Meal Preparation	Laundry Work[b]
No. of children under 1 year	22.73 (4.67)	
No. of children 1-year-old	20.39 (4.13)	
No. of children 2–5	18.28 (7.58)	
No. of children 6–11	16.14 (8.52)	
No. of children 12–18	6.52 (2.78)	
Wife's education	.02 (.01)	−2.29 (1.96)
Husband's education	1.16 (.70)	−.75 (−.81)
Age of wife	.92 (3.95)	
Husband's time input	−.51 (−5.74)	
Rooms per capita	1.61 (.40)	.35 (.23)
Nonautomatic washer used[a]		38.53 (7.8)
Dryer used[a]		−9.69 (−4.51)
Washer-dryer used[a]		8.01 (.95)
No. of loads of wash		13.69 (21.7)
Laundromat used[a]		6.59 (1.89)
C	94.72 (7.61)	11.92 (2.40)
R^2	.14	.46

[a] Minutes over two days.

[b] Dummy variables equal to 1 if statement is true, zero otherwise. The omitted categories are automatic washer and line drying.

Note: t-values in parentheses.

educated women have more and better quality capital goods, but even with these factors held constant in the regression, education still serves to decrease time inputs to laundry work.

In Table 3, we see time inputs to physical care of children are closely related to the number and ages of children. While each child under one year requires an additional 208 minutes of care over a two-day period from his mother, a one-year-old adds 112 minutes, and each two- to five-year-old adds only 26 minutes. Older children may reduce the demands on the mother's time, presumably by helping with the feeding and dressing of younger siblings.

The productivity effect, measured by the wife's education, is positive but not very significant. The weak effect of education on time devoted to child care probably reflects the outcome of two opposing forces: a negative substitution effect, whereby education raises the wife's potential market wage, pulling her into the labor market; and a positive substitution effect whereby education increases home productivity and orients the wife toward home production.

The income proxies, husband's education and rooms per capita, are not significantly different from zero. The latter has the wrong sign possibly because greater numbers of rooms indicate not only greater income but also greater demand for time in such activities as cleaning. Women who prefer physical care spend more time at it than women who rate it less highly.[5]

While husband's time is a substitute for the wife's in meal preparation, this is not the case for physical care. The significant positive coefficient on husband's time inputs indicates that, for each ten minutes the husband spends in physical care, the wife puts in an additional four minutes. This is not merely an indication of the family's tastes (since these are controlled for by the "preference" variable), but may indicate true complementarity—increased inputs of husband's time increasing the marginal productivity of the wife's time inputs. Time spent by others in the care of children had no significant effect on the mother's time inputs.

The sample was next divided into a high-education group including 493 women with some college training, and a low-

---

[5] See Table 1 for a definition of the preference variable.

TABLE 3—REGRESSIONS ON TIME INPUTS TO CHILD CARE

	Physical Care			Other Care		
Schooling Group:	All	High	Low	All	High	Low
No. of children under 1	207.6 (26.24)	214.8 (19.02)	204.9 (19.97)	23.42 (3.50)	6.62 (.71)	46.84 (4.84)
No. of children 1 year	111.6 (14.52)	114.88 (9.99)	114.81 (11.63)	30.53 (4.52)	36.70 (3.71)	29.78 (3.19)
No. of children 2–5	26.34 (6.92)	19.81 (3.59)	36.55 (7.55)	18.03 (5.40)	15.13 (3.26)	21.96 (4.55)
No. of children 6–11	−4.91 (−1.25)	−4.03 (−.96)	−.48 (−.13)	15.47 (5.84)	17.99 (4.77)	14.14 (3.78)
No. of children 12–18	−6.99 (−1.88)	−5.00 (−.93)	−4.41 (−.99)	.91 (.28)	−.73 (−.15)	3.27 (.74)
Wife's education	2.96 (.91)	6.90 (.97)	5.70 (.85)	.35 (.12)	−5.45 (−.95)	−2.95 (−.48)
Husband's education	1.62 (.64)	.60 (.05)	2.08 (.66)	3.48 (1.57)	−2.19 (−.65)	−3.54 (−1.20)
Age of wife	1.45 (1.82)	.77 (.57)	2.03 (2.08)	.32 (.46)	−.79 (−.72)	1.23 (1.36)
Husband's time input	.389 (4.18)	.53 (3.91)	.27 (2.05)	.287 (7.46)	.25 (4.73)	.344 (6.24)
Rooms per capita	−11.84 (−1.41)	.25 (.34)	−.42 (−.66)	6.17 (.84)	3.08 (.29)	9.47 (.93)
Husband's age	−1.65 (−2.21)	−1.63 (−1.31)	−1.74 (−1.87)	−.68 (−1.05)	.22 (.22)	−1.44 (−1.67)
Care by others	−.005 (−.05)	.20 (1.18)	−.14 (−1.07)	−.04 (−1.74)	.010 (.33)	−.123 (−3.40)
Preference	3.94 (3.03)	3.96 (1.99)	3.83 (2.20)	.55 (.38)	−.78 (−.36)	1.14 (.57)
C	68.92 (2.45)	71.52 (1.70)	46.53 (1.17)	53.27 (2.04)	84.01 (2.11)	32.28 (.55)
R^2	.65	.68	.63	.15	.16	.18

Note: t-values in parentheses.

education group including 591 women who had not gone beyond high school. Time input regressions run on the two subsamples indicate that husband's time is more complementary for the high-education group. Time inputs by others—older children and adults other than the parents—seem to act as substitutes for own time in the low-education group, but not for those with more schooling.

The next three equations deal with time inputs to other care—which includes time spent in social and educational activities with children. Time inputs are again clearly related to the number and ages of children. Again the husband's time inputs are complementary to the wife's time, while time inputs by others are only weak substitutes for the wife's time. For each 100 minutes of care by others, the wife reduces her own time inputs by only three minutes. The positive coefficient on husband's education may reflect a positive income effect, but rooms per capita, the second income proxy, is not significantly related to time inputs.

Time inputs by others have a negative impact on time spent with children by less-schooled mothers (which is significant at the 1 percent level), but they have no effect on time inputs by mothers with more schooling. That is, mothers with less schooling act as if time inputs by older children and other adults are a better substitute for their own time than do the more-educated mothers.

The substitutes—baby sitters, grandmothers, other children over six—are more similar in education and ability to the mothers with little schooling. However, if education increases the productivity of time in child care, more-educated women would find these other workers relatively unsatisfactory substitutes. In fact, mothers in the high-education group spend the same amount of time in child care whether or not other workers also care for their children.

Consistent with these results is Stafford and Hill's finding (1973b) that in high-status families (defined by husband's education or occupation) time inputs to preschool children are invariant with the number of other children in the family, while, in low-status families, mothers reduce their time inputs to preschool children as family size increases. This behavior on the part of more-educated mothers, in addition to their greater price of time, increases the "cost" to them of children and acts to reduce their demand for numbers of children. Further, since the mother's time inputs will not be reduced when older children care for their younger siblings, there is less incentive for wide spacing of children on the part of more-educated mothers. In fact, Sue Ross has shown that more-educated mothers have shorter intervals between births.

The fact that more-educated women spend more of their own time in child care in spite of the higher price of their time is due partly to the low substitution elasticity with other inputs, as contrasted with two other household activities. Educated women tend to be in high-income families and have a greater productivity of time in child care relative to other household activities; both factors raise the demand for time in child care (Leibowitz, 1972). Despite the income proxies' lack of statistical significance, there is inferential evidence for the family income effect, since the fact that husbands and other adults in families where the wife is highly educated devote more time to child care can scarcely be explained by the greater productivity of more-educated women's time.

The historical evidence also supports the hypothesis that the income elasticity of demand for own time in child care is high. Vanek finds in samples ranging back to 1926 that higher *SES* women spend more time in child care, and that "contemporary nonemployed women spend about ten hours a week more in family care than women in earlier samples." That the income elasticity of child care exceeds that for other housework can also be inferred from Vanek's finding that contemporary women lower their standards for housework when child-care demands are highest.

The considerable time devoted to child care by better-educated parents represents substantial investment in the human capital of their children. The finding that time investments are greater for more-educated mothers helps to explain the observed positive correlation between children's achievement and their mother's schooling (J. Kagan and H. A. Moss). Indeed, evidence is beginning to accumulate that these investments affect not only the early achievement of children but also their ultimate level of schooling (Leibowitz, 1974b).

The time allocations of more-educated women in the 1920's foreshadowed the allocation decisions of all women in the 1960's. If the contemporary sample also forecasts the future, we can cautiously predict that, when the average woman has more schooling and higher income than today, she will indeed spend a greater proportion of her lifetime in the labor force. Although she will probably spend less of her increasingly valuable time in most home production activities, she is likely to spend even more of it with her children.

REFERENCES

L. **Benham**, "Benefits of Women's Education Within Marriage," *J. Polit. Econ.*, forthcoming.

Y. **Ben-Porath**, "The Production of Human Capital Over Time," in W. L. Hansen, ed., *Education, Income and Human Capital*, New York 1970.

W. G. **Bowen** and T. A. **Finegan**, *The Economics of Labor Force Participation*, Princeton 1969.

G. G. **Cain**, *Married Women in the Labor Force*, Chicago 1966.

M. **Grossman**, "The Correlation Between Health and Schooling," Nat. Bur. of Econ. Res., mimeo., 1973.

J. **Kagan** and H. A. **Moss**, "Parental Correlates of Child's I.Q. and Height," *Child Development*, *30*, 325–32.

A. **Leibowitz**, "Women's Allocation of Time to Market and Non-Market Activities: Differences by Education," unpub. Ph.D. diss., Columbia Univ. 1972.

———, "Education and the Allocation of Women's Time," in F. T. Juster, ed., *Education, Income and Human Behavior*, New York 1974a.

———, "Home Investments in Children," *J. Polit. Econ.*, Part II, Mar.–Apr. 1974b.

S. L. **Manning**, "Time Use in Household Tasks in Indiana Families," Purdue Univ. Agricultural Experiment Station, res. bul. 837, Jan. 1968.

R. T. **Michael**, "Education and the Derived Demand for Children," *J. Polit. Econ.*, Part II, Mar.–Apr. 1973, *81*, 128–64.

J. **Mincer**, "Labor Force Participation of Married Women: A Study of Labor Supply," in A Conference of the Universities, Nat. Bur. of Econ. Res., *Aspects of Labor Economics*, Princeton 1962.

S. **Ross**, "Timing and Spacing of Births," Nat. Bur. of Econ. Res., mimeo., 1972.

F. P. **Stafford** and C. R. **Hill**, "Allocation of Time to Preschool Children and Educational Opportunity," mimeo., 1973a.

——— and ———, "Time Inputs to Children by Parents' Education, Occupation, Religion and Sex of Child," Univ. of Michigan, mimeo., 1973b.

J. **Vanek**, "Keeping Busy: Time Spent in Housework, United States, 1920–1970," unpub. Ph.D. diss., Univ. of Michigan 1973.

K. **Walker**, Use-of-time-Research Project, "Definition of Household Activities," N.Y. State Coll., of Human Ecology, Cornell Univ. 1967.

R. **Willis**, "A New Approach to the Economic Theory of Fertility Behavior," *J. Polit. Econ.*, Part II, Mar.–Apr. 1973, *81*, 14–64.

The Effects of Government Expenditures and Taxes on Female Labor

By Michael J. Boskin*

It is well known that government expenditures and taxes affect virtually all economic activity, including the allocation of work effort between the market and the home, the accumulation of human and physical capital, marriage, divorce and fertility behavior, consumption decisions, retirements and bequests as well as the timing over the life cycle of all of these activities.[1] It is not commonly known, however, that for a number of important activities the public budget affects male and female labor differently, both because of numerous special provisions in the tax laws and because males and females (or husbands and wives) sometimes respond differently to the same incentives. Examples of the former include income splitting under the income tax and the spouse's benefit under the social security system. Examples of the latter include the effects of taxes on market labor supply and occupational choice. The purpose of the present paper is to present an overview of some of the most important of these effects, together with some preliminary empirical evidence on their absolute magnitude and relative importance.

My method of presentation is to discuss the major economic decisions and the effects of government expenditures and taxes upon them roughly in life-cycle order and in the context of a model of expected utility maximization. While this approach facilitates the discussion, it is important to note that these resource allocation decisions are made (and continually revised) simultaneously and that the allocative and distributive effects of any particular tax device or expenditure program depend upon the other taxes and expenditures in the system. At the general and nontechnical level of this presentation, such general equilibrium effects, which form a large part of my current research, are mostly ignored. However, some preliminary indications of the combined effects of several taxes are presented where they seem important and are known empirically.

I. Conceptual Framework

We assume that individuals behave so as to maximize the expected present value of utility over their lifetime. Utility at any point in time is a function of the consumption of market goods and services and household goods and services and of bequests made to others. Market goods and services are purchased directly, whereas household goods and services must be produced with purchased capital and the time of the individual (and perhaps the time of other individuals). The individual starts "life" (say at age 18) with endowments of physical and human capital. Human capital is produced with inputs of time; the individual may borrow and lend nonhuman assets at a given rate of return. The budget constraint of the individual requires that the present value of consumption (including the value of time used for work in the home or human capital accumulation) not exceed "full wealth" (human plus nonhuman) net of taxes and transfers. The fixed time available per period is allocated among market and household work and human capital accumulation.

---

* Stanford University. This paper presents a brief, nontechnical overview of research in progress. I am indebted to Victor Fuchs for encouraging me to pursue this topic and for valuable comments, and to Mark R. Killingsworth for several excellent suggestions.

[1] Of course, government expenditures and taxes affect the cyclical performance of the economy; while the effects of such cyclical behavior on female labor are not unimportant, we shall not deal with them here.

From such a formulation, given specific utility and production functions, we may derive the optimal levels and time patterns of the individual's market labor supply, work in the home, human and physical capital accumulation and bequest.[2] These in turn depend upon age and the levels and time patterns of real after-tax wage and rental rates, rates of return to human and nonhuman capital, the price of market output, the risk aversion of the individual and the initial holdings of human and nonhuman wealth. Hence, various tax devices and government expenditures which affect these parameters in turn affect all of the activities mentioned above. While the interaction of all of the effects of. a tax- or expenditure-induced change in any of these parameters is quite complex, we note the following effects and their probable results.

Wage effects. Recent work (see Boskin, 1973a, and Robert Hall) suggests that the wage elasticity of wives' market labor supply is positive and much larger than that of husbands. Thus, tax-induced reductions in the after-tax wage directly reduce female labor supply. This is reinforced by two indirect effects: the reduction in time spent in the market reduces learning-by-doing (i.e., reduces gains in market productivity generated during time devoted to market work), and/or accelerates the depreciation of market skills relative to home skills (see Jacob Mincer and Solomon Polachek). Both indirect effects reduce the wage, and hence female market labor supply, still further.

Tax-induced reductions in the real market wage also directly affect the return to, and thus the actual amount of, investments in human capital. This phenomenon has received relatively little attention, so not much is known about sex differences in the rate-of-return or wage elasticity of such investment. However, my own results on occupational choice (1974) suggest that this particular market-specific human capital investment is highly wage elastic. Further, decreases in the posttax wage make it relatively more desirable to specialize in types of human capital accumulation specifically useful in household, as opposed to market, production, and it is females who are more likely to spend a greater relative percentage of their time in household production activities than men.

Wage changes also produce two types of "cross" effects. The first is the familiar cross-wage effect on the market labor supply of other household members, which may be decomposed into cross-substitution and income effects. While economic theory implies that the cross-*substitution* effects for different members are equal, it places no restrictions on the *total* cross-wage effects. The scanty available empirical evidence (see Boskin, 1973a) suggests that wives' market labor supply is more sensitive to changes in the husbands' wage than vice versa and that the effect is negative. Thus, to a limited extent, the negative cross-wage effect offsets the differential own-wage effect.

The second "cross" effect of wage changes relates to nonhuman capital accumulation. The wage-induced decrease in human capital accumulation produces an income and cross-substitution effect on nonhuman capital accumulation. The former operates to decrease, and the latter to increase, nonhuman capital accumulation. There is very little empirical evidence on these effects.

---

[2] One may also incorporate marriage, divorce and fertility behavior into the analysis by examining the efficient production of household output. See Gary Becker.

Finally, we note that the general equilibrium consequences of a tax-induced change in after-tax market wage rates depend upon the elasticities of factor-substitution in market and household production, the elasticity of demand for market output, factor intensities, and factor shares and that such a change may entail changes in the price of capital and the shadow price of household labor. These in turn have their own effects, and it is the net effect of all of these changes, i.e., the reduced form of the general equilibrium system, in which we are interested.

Nonwage income effects. We generally argue that goods are normal—that, *ceteris paribus*, as income increases, more of each good is demanded. Hence, a tax-induced decrease in nonwage income tends to decrease the demand for leisure—time spent out of the market—and the accumulation of both market- and home-specific human capital. It does appear, however, that at least in the case of market labor supply, income effects differ slightly for husbands and wives, but are quite small (see Boskin, 1973a).

Rental price of nonhuman capital effects. The effects of tax changes on the rental price of nonhuman capital, and hence on its rate of return and on the allocation of time, are a bit more subtle than the direct effects of wage changes. Factor-substitution, factor-intensity and demand considerations influence the allocation of the capital stock—also the labor force—between the market and the home; this in turn may alter the wage rate.

A decline in the rate of return to capital may induce a decrease in savings which in turn entails a cross-substitution effect tending to increase, and an income effect tending to decrease, human capital accumulation. Some empirical evidence, however, suggests that private savings are quite insensitive to changes in the rate of return (see Paul David and John Scadding).

Effects of uncertainty. The responses of market labor supply, human capital accumulation and savings to changes in risk are again quite complex and depend upon whether nonwage income, the rate of return to capital, or wage rates are uncertain. For example, a risk-averse individual will hedge against an increase in income uncertainty by saving more and working more in the market. When factor returns are uncertain, however, the net effect of an increase in risk is not analytically determinate, as it is composed of a substitution effect away from the more risky factor and the income effect mentioned above working in the opposite direction. There is also a stochastic cross-substitution effect about which nothing can be said *a priori*.[3] Unfortunately, empirical evidence on male-female differences—if any—in behavior toward risk is not available.

How, then, do specific tax and expenditure programs fit into this general picture? Which effects do they have? Are they large enough to be important? What are the net effects of the tax and expenditure system combined? To this we now turn.

II. Effects of Taxes

Of the many effects of taxes on female labor, the most important are probably those of the individual income and payroll taxes on market work effort and human capital accumulation. The wage effect of these tax devices depends upon their marginal, and the income effect upon their average, rates. Federal, and state and local, individual income taxes and (both parts of) the payroll tax substantially reduce real after-tax wage rates even at modest levels of gross income. Even if these taxes were levied at flat rates on all income, they would still drive female labor out of the market into the home, since household labor is untaxed and, as noted above, the market labor supply of wives is more wage elastic than that of husbands. This effect is reinforced by a variety of special provisions of these two taxes.

---

[3] It might be useful to point out that uncertainty and the incentive for family formation are directly related. While the variance of the sum of the incomes of husbands and wives will increase or decrease depending upon whether their incomes are positively or negatively correlated, the variance of per capita income will decline so long as the two incomes are not perfectly correlated. Hence family formation (or any other type of income sharing) can be used as a hedge against the vagaries of labor and financial markets.

First, the progressive rate structure combines with the income splitting provision to drive the marginal tax rate on the *first* dollar earned by the second earner in the family—usually the wife[4]—up to the rate on the *last* dollar earned by the first earner (see Boskin, 1973b). This heavy taxation of market, as opposed to household, labor tends to reduce both market labor supply and market-oriented human capital accumulation by (present or prospective) wives, since it reduces the after-tax wage and the rate of return to such investment by a percentage no less than the marginal tax rate. Since this rate of return also depends upon the extent and timing of market work, the market labor supply disincentive reinforces, in sort of a vicious cycle, this disincentive to accumulate human capital. It is further reinforced by any reductions in learning-by-doing or by relatively more rapid depreciation of market skills in the home than in the market. Worse yet, husbands reach their peak earnings, and hence marginal tax rates, at the time many women consider reentering the labor force after rearing their children.

Probably less important than the income splitting provision are the dependent exemptions and certain deductions which tend to lower the price of inputs into household production. The exemptions and the deduction for employment-related expenses for the care of a dependent reduce the cost of children, which are complementary to female labor, whereas the deduction of mortgage interest and the exclusion of the imputed income of owner-occupied homes reduce the cost of household capital, which may substitute for female labor. The net effect on household labor again depends upon whether the incentive to substitute capital for labor in household production is as strong as the incentive to produce more household output.

Income averaging, income splitting and the progressive rate structure all combine to reduce risk and hence to offset the relatively larger market labor supply and capital accumulation generated under income uncertainty; income splitting also reinforces the income uncertainty incentive to marry. The empirical importance of changes in risk on market labor supply and capital accumulation is unknown.

Finally, the payroll tax places a ceiling on taxable earnings. This converts the tax's wage effect into an income effect for persons with earnings above the ceiling; it thus tends to increase market labor supply relative to a tax of equal yield levied at lower rates over all earnings and to reduce the disincentive to accumulate human capital caused by such a tax. It also induces employers to substitute skilled for unskilled, and avoid part-time, workers, which again tends to shift female labor into the home.

The effects of taxes on capital on the allocation of work between the market and the home and on capital accumulation by type are both less clear analytically and more difficult to establish empirically than the analogous effects of taxes on earnings. While taxes on capital in the market, such as the corporation income tax, the relevant part of the individual income tax and the commercial property tax, induce a substitution of labor for capital in the production of any given level of market output, they also tend to raise the price of market output and induce a substitution of household output in consumption. The net outcome on the allocation of labor between the market and the home depends upon the relative weights of factor substitution, demand and factor intensity effects; my previous work (Boskin, 1973c) suggests that the relatively heavier taxation of market than household capital is likely to drive labor—and it is usually wives that are most sensitive in this movement—into the home. The effect of such taxes on human capital accumulation is still more subtle and analytically indeterminate. The risk effects of income and capital gains taxation on market labor supply and human capital accumulation again are quite complex, depending in part upon the degree of risk aversion.

---

[4] This may be due to a comparative advantage in the home for female labor, role discrimination, or cultural and psychological factors. We eschew any attempt at examining this important question here; what is important for our purposes is that, for whatever reason, it is usually female labor which moves between market and household work.

Finally, the tax system also affects bequests. If bequests are normal goods, the income effect of the tax system reduces bequests. The estate and gift tax creates an incentive to bequeath in the form of human capital (paying for an education); it also contains a spouse's exemption. There is no capital gains tax on assets passed on at death and this influences the pattern of bequests over the life cycle. It is an open question as to whether males and females respond differently to changes in endowments of human and nonhuman capital.

III. Effects of Government Expenditures and Transfers

When the public sector spends tax proceeds differently than the private sector would have spent them in the absence of the taxation, it induces a set of factor and product price changes and a reallocation of resources. Since the public sector probably has a higher marginal propensity to consume market output than does the private sector and since market output is relatively labor-intensive (see Boskin, 1973c), the net effect is to increase the rate of return to labor relative to capital, to drive up the relative price of market output, and probably to attract labor into the market. This effect is reinforced by the usually labor-intensive nature of direct public production of goods such as education.

Most transfer programs tend to work in the opposite direction. The most obvious example is the welfare system. For example, the income guarantee and the extremely large implicit tax rates in Aid to Families with Dependent Children create a strong market labor supply disincentive at the lower end of the income scale. It also creates an unfortunate incentive for families to split up. Finally, it operates, as does the unemployment insurance system, to reduce income uncertainty and produces the effects noted above.

The structure of social security benefits creates a number of conflicting incentives. The benefit formula creates a strong incentive to participate in covered employment at least long enough to qualify for benefits. The system also may partially displace private capital accumulation. However, the spouse's benefit provision entitles a wife (or husband) to one-half of her (his) spouse's benefit. Since the benefit is the larger of the "own" benefit and one-half of the spouse's benefit, wives have little incentive to become covered under social security unless they expect an "own" entitlement substantially in excess of one-half of their husband's entitlement. Further, the widow's benefit acts as savings insurance for married persons. The benefit structure still retains a substantial implicit tax on earnings (beyond $2,400 a year) for social security beneficiaries.

Finally, the extensive system of publicly financed education transfers resources between generations, i.e., effects tax-free social bequests of human capital. This bequest may be larger than would be forthcoming, due to external benefits, from the private sector and probably partially offsets the effect of the tax system on human capital accumulation.

IV. Conclusion

The net effect of the tax system is clearly to drive female labor out of the market into the home and to reduce the (market-oriented) human capital accumulation of women relative to men. These effects may be slightly offset by government expenditures on goods and services, but they are reinforced by existing transfer programs.

The net effect of the public budget is thus to introduce a substantial misallocation of resources.[5] With the social marginal product given gross-of-tax and private decisions made net-of-tax, society suffers an unnecessary cost in driving a wedge between the marginal social product of labor in the market and the home. I have elsewhere (Boskin, 1973c) estimated a rough lower bound on this annual welfare loss to be $20 billion (although a minor part of this is due to the misallocation of the capital stock). To this must be added the annual benefits which society foregoes due to underinvestment in human capital. These tax-induced misallocations of resources are largest for female labor. This deadweight loss to society, it seems to me, is too large to ignore in formulating tax and expenditure policy.

REFERENCES

G. Becker, "A Theory of Marriage: Part I," *J. Polit. Econ.*, Aug. 1973, *81*.

M. Boskin, "The Economics of the Labor Supply," in G. Cain and H. Watts, eds., *Income Maintenance and Labor Supply*, Chicago 1973a.

———, "Optimal Tax Treatment of the Family," memo. 143, Cent. for Res. in Econ. Growth, Stanford 1973b.

———, "Efficiency and Equity Aspects of the Differential Tax Treatment of Market and Household Economic Activity," memo. 149, Cent. for Res. in Econ. Growth, Stanford 1973c.

———, "A Conditional Logit Model of Occupational Choice," *J. Polit. Econ.*, Mar.–Apr. 1974, *82*.

P. David and J. Scadding, "Private Saving: Ultra-Rationality, Aggregation and Denison's Law," *J. Polit. Econ.*, Mar.–Apr. 1974, *82*.

R. Hall, "Wages, Income and Hours of Work in the U.S. Labor Force," in G. Cain and H. W. Watts, eds., *Income Maintenance and Labor Supply*, Chicago 1973.

J. Mincer and S. Polachek, "Family Investments in Human Capital: Female Earnings," *J. Polit. Econ.*, Mar.–Apr. 1974, *82*.

---

[5] I assume that these distortions are not justified on other grounds.

PRODUCTION AND THE AGGREGATE DEMAND FUNCTION FOR LABOR

The Elasticity of Substitution and Cyclical Behavior of Productivity, Wages, and Labor's Share

By THOMAS J. SARGENT AND NEIL WALLACE*

It is commonly asserted that the behavior of time series on output, capital and labor implies almost nothing about the value of the elasticity of substitution between capital and labor, σ.[1] In what follows we show that in Robert E. Lucas's framework, the behavior of time series on output, capital and labor, on the contrary, has very sharp implications about the value of σ.

Lucas was interested in reconciling diminishing returns to labor with two longstanding and paradoxical empirical observations: the procyclical behavior of output per manhour (reflected in, for example, time-series estimates of the Cobb-Douglas production function showing increasing returns to manhours);[2] and the Dunlop-Tarshis observation of the failure of average hourly real earnings to move countercyclically, in apparent contradiction of the postulates of both Keynes and the Classics.[3] Lucas's model effects a reconciliation by noting that the labor-capital ratio varies systematically over the work day, leading to systematic differences over the day in the marginal productivity of labor. His theory explains the paradoxical observations as resulting from the way data have been averaged over the day. The model provides one indication of the probable returns to taking seriously Nicholas Georgescu-Roegen's call (p. 9) for a more careful treatment of the length of the work day in the analysis of production. Here we point out that Lucas's ingenious explanation of the paradoxical cyclical behavior of both output per manhour and average hourly real earnings seems to require that σ be (much) less than unity.

In the usual model, if σ is less than unity, the distributive share of labor is a countercyclical variable. But in the post-World War II period, labor's share has failed to display a pronounced countercyclical pattern.[4] Here it will be shown that in the con-

---

* Department of Economics, University of Minnesota, and consultants, Federal Reserve Bank of Minneapolis. This paper was financed by the Federal Reserve Bank of Minneapolis, which is not, however, responsible for the opinion expressed here.

[1] For example, see Robert E. Hall and Dale Jorgenson.

[2] Estimates of the Cobb-Douglas production function with constant returns to scale for time series typically show constant or increasing returns to man-hours and approximately zero or negative returns to capital. See Ronald Bodkin and Lawrence Klein and Lucas. These empirical results are widely ignored; for example, see Franklin M. Fisher. Fisher proceeds as though the time series estimates of the constant returns Cobb-Douglas function yield a capital elasticity of about .25.

[3] For a summary of the empirical literature on the cyclical behavior of the average real wage bill per manhour, see Bodkin.

[4] It is frequently asserted, largely on the basis of observations drawn from the Great Depression, that labor's distributive share is countercyclical. However,

text of Lucas's model, labor's distributive share need not move countercyclically and can even move procyclically if σ is less than unity.

Following Lucas, we suppose that during the time period between t and t+1 production at instant t+s is governed by the instantaneous aggregate production function

$$(1)\qquad y(\mathrm{t}+s)=f[n(\mathrm{t}+s),k(\mathrm{t})]$$
$$\mathrm{t}=0,1,2,3,\ldots;\quad 0\leq s<1$$
$$f_k,f_n\geq 0;\qquad f_{nn},f_{kk}<0$$

where $y(\mathrm{t}+s)$ is the rate of real output per unit of time at instant t+s, $n(\mathrm{t}+s)$ is the stock of employment at instant t+s, and $k(\mathrm{t})$ is the capital stock, which is assumed to be constant over the time interval between t and t+1. We assume that $f(n, k)$ is linearly homogeneous in n and k. The period between t and t+1 can be thought of as one day, being divided into several work shifts. All firms in the industry or economy are assumed to be identical, the rate of production in each firm being described by equation (1).

Suppose now that the representative firm is faced with a real wage $w(\mathrm{t}+s)$ at which it can hire all the labor it wishes at instant t+s. The real wage is assumed to vary over both t and s in the particular fashion described by

$$(2)\qquad w(\mathrm{t}+s)=w_0(\mathrm{t})\phi(s).$$

The real wage at instant t+s equals the product of the "straight-time" real wage $w_0(\mathrm{t})$ and the premium $\phi(s)$. The premium function is assumed *not* to change from day to day, and so does not depend on t. We can let $\phi(s)$ take on the value unity for straight-time work, and take on a value greater than unity for, e.g., the graveyard shift or "overtime."

The firm maximizes daily profits with respect to employment at each instant during the day. It accomplishes this by equating the marginal product of labor with the real wage at each instant:

$$(3)\qquad \partial f[n(\mathrm{t}+s),k(\mathrm{t})]/\partial n(\mathrm{t}+s)=w_0(\mathrm{t})\phi(s)$$

Since $f_{nn}<0$, employment at instant t+s varies inversely with $\phi(s)$. That is, the firm hires fewer workers on "later" shifts because it must pay higher real wages for work during those hours. At the margin, then, workers on "later" shifts are more productive than those on "earlier" ones because they work with more capital per man.

Lucas pointed out that where employment and output are determined by (1), (2), and (3), "daily" output per manhour may or may not vary inversely with the manhours-capital ratio, even though diminishing returns to labor prevail at each instant. That is, even though instantaneous output is related to instantaneous employment via equation (1), it is not in general true that "daily" output, $\int_0^1 y(\mathrm{t}+s)ds$, is related to "daily" manhours, $\int_0^1 n(\mathrm{t}+s)ds$, by

---

the data indicate that labor's share has not behaved strongly countercyclically in the United States during the post-World War II period. Regressing two measures of labor's share against time and the unemployment rate (Un_t) gave the following results for quarterly data for the United States from 1948I–1971IV:

$$s_{1\mathrm{t}}=\underset{(.0041)}{.4961}+\underset{(.000034)}{.000089\mathrm{t}}-\underset{(.00080)}{.00028\,Un_{\mathrm{t}}}$$
$$R_A^2=.0494,\ d.w.=.16$$

$$s_{2\mathrm{t}}=\underset{(.0040)}{.5571}-\underset{(.000033)}{.000159\mathrm{t}}+\underset{(.00078)}{.00024\,Un_{\mathrm{t}}}$$
$$R_A^2=.1850,\ d.w.=.17$$

Standard errors are in parentheses. Here s_1 is compensation of employees minus the government wage bill divided by *GNP* minus the government wage bill; s_2 is compensation of employees plus .65 times income of unincorporated businesses minus the government wage bill divided by *GNP* minus the government wage bill. Neither share concept moved strongly with the unemployment rate. For discussions to the effect that labor's share has failed to move countercyclically during the postwar period in a number of Western European countries, see Jean Marchal and Bernard Ducros (pp. xxvii–xxviii, p. 74).

$$\int_0^1 y(\mathrm{t}+s)ds = f\left[\int n(\mathrm{t}+s)ds,\ k(\mathrm{t})\right].$$

(Since integration is over [0, 1] throughout the paper, we omit the limits.) Diminishing marginal returns to instantaneous employment is simply not sufficient to imply diminishing returns of daily output with respect to daily manhours. Furthermore, even though the instantaneous real wage obeys (3), and so varies inversely with the instantaneous employment-capital ratio, there is no implication that average hourly real earnings over the day varies inversely with the daily manhours-capital ratio. As a result, the paradoxical empirical observations that average hourly real earnings and output per man-hour are not countercyclical (i.e., don't vary inversely with the manhours-capital ratio) need not be interpreted as contradicting the hypothesis of diminishing returns to labor.

But those empirical observations can be rationalized within Lucas's framework only if the elasticity of substitution between capital and labor is (much) less than unity. To begin, suppose on the contrary that the production function is Cobb-Douglas so that (1) has the particular form

$$\text{(4)}\qquad y(\mathrm{t}+s) = An(\mathrm{t}+s)^{\beta}k(\mathrm{t})^{\alpha};\quad 1 > \alpha > 0,\quad \beta = 1-\alpha$$

Solving for the level of employment at t+s gives, from the marginal product condition,

$$\text{(5)}\quad n(\mathrm{t}+s) = [w_0(\mathrm{t})\phi(s)/\beta Ak(\mathrm{t})^{\alpha}]^{[1/(\beta-1)]}.$$

Upon integrating $n(\mathrm{t}+s)$ and $n(\mathrm{t}+s)^{\beta}$ over the day, we have

$$\text{(6)}\quad \int n(\mathrm{t}+s)^{\beta}ds = \left[\int n(\mathrm{t}+s)ds\right]^{\beta}(A'/A)$$

where A', a constant, is given by

$$A' = A\int \phi(s)^{[\beta/(\beta-1)]}ds \bigg/ \left[\int \phi(s)^{[1/(\beta-1)]}ds\right]^{\beta}.$$

Thus, daily output Y(t), which is found by integrating (4) over the day, is

$$\text{(7)}\quad Y(\mathrm{t}) = Ak(\mathrm{t})^{\alpha}\int n(\mathrm{t}+s)^{\beta}ds = A'k(\mathrm{t})^{\alpha}\left[\int n(\mathrm{t}+s)ds\right]^{\beta},$$

a loglinear function of capital and daily man-hours, with the same capital and labor elasticities that characterize the instantaneous production function. Consequently, if diminishing returns to labor characterize the instantaneous production function ($\beta<1$), then "daily" output will appear to display diminishing returns with respect to "daily" manhours.

The average hourly real wage is given by

$$\bar{w}(\mathrm{t}) = \int w(\mathrm{t}+s)n(\mathrm{t}+s)ds \bigg/ \int n(\mathrm{t}+s)ds,$$

i.e., the daily real wage bill divided by daily manhours. Substituting for $n(\mathrm{t}+s)$ from (5) gives

$$\text{(8)}\quad \bar{w}(\mathrm{t}) = w_0(\mathrm{t})\int \phi(s)^{[\beta/(\beta-1)]}ds \bigg/ \int \phi(s)^{[1/(\beta-1)]}ds,$$

a constant times the straight-time real wage. Thus the movement over time of average hourly real earnings provides a reliable indication of the movement of the straight-time real wage. If the instantaneous production function is Cobb-Douglas, then (a) Lucas's model predicts that both average hourly real earnings and output per manhour will be countercyclical, i.e., they will vary inversely with the man-hours-capital ratio; and (b) Lucas's model cannot rationalize procyclical behavior of labor's distributive share. Labor's share, ξ, is given by

$$\xi = \int w(\mathrm{t}+s)n(\mathrm{t}+s)ds \Big/ \int f[n(\mathrm{t}+s), k(\mathrm{t})]ds,$$

which in the Cobb-Douglas case is simply β, a constant.

To show what is required for Lucas's model to be able to explain the empirical observations, we begin by calculating the elasticities of daily output per manhour and the average hourly real wage with respect to the straight-time real wage, $\epsilon_{Y/MH,w}$ and $\epsilon_{\bar{w},w}$, respectively. Daily output per manhour is

$$Y(\mathrm{t})/MH(\mathrm{t}) = \int f[n(\mathrm{t}+s), k(\mathrm{t})]ds \Big/ \int n(\mathrm{t}+s)ds,$$

so that we have

$$(9) \quad \epsilon_{Y/MH,w} = [w_0(\mathrm{t})/Y(\mathrm{t})] \int \phi(s) \cdot [\partial n(\mathrm{t}+s)/\partial w(\mathrm{t}+s)] \cdot [(\partial f/\partial n) - Y(\mathrm{t})/MH(\mathrm{t})]ds.$$

From the definition of $\bar{w}$,

$$(10) \quad \epsilon_{\bar{w},w} = 1 + [w_0(\mathrm{t})/\bar{w}(\mathrm{t})MH(\mathrm{t})] \int \phi(s) \cdot [\partial n(\mathrm{t}+s)/\partial w(\mathrm{t}+s)] \cdot [w(\mathrm{t}+s) - \bar{w}(\mathrm{t})]ds.$$

Thus, it will help make these elasticities negative if the absolute value of $\partial n(\mathrm{t}+s)/\partial w(\mathrm{t}+s)$ varies directly with $w(\mathrm{t}+s)$; that is, if employment at instants with marginal products of employment exceeding output per manhour ($\partial f/\partial n > Y(\mathrm{t})/MH(\mathrm{t})$, high-wage instants) is more (inversely) responsive to the real wage rate than is employment at instants for which $\partial f/\partial n < Y(\mathrm{t})/MH(\mathrm{t})$, low-wage instants. Of course, for there to be instants at which $\partial f/\partial n > Y(\mathrm{t})/MH(\mathrm{t})$, the premium function must be appropriately shaped.

We have already seen that for any premium schedule Lucas's model predicts that daily output per manhour and the average hourly real wage both vary directly with the real wage if the instantaneous production function is Cobb-Douglas. For the Cobb-Douglas function with constant returns to scale, the logarithm of the marginal product of employment is a linear function of the logarithm of the employment-capital ratio as depicted in Figure 1. Since $\phi(s)$ is constant over time, employment at instant $\mathrm{t}+s$ is a constant multiple of employment at instant t. Thus, the proportions of employment at various instants over the day remain unchanged as the straight-time real wage changes; see (5). To explain the observations on output per man-hour and average hourly earnings, Lucas's model seems to require an instantaneous production function that predicts that a decrease in the straight-time real wage produces a sizable increase in the proportion of workers employed at high productivity instants.

Suppose then that the instantaneous production function is the constant elasticity of substitution (*CES*) function

$$y(\mathrm{t}+s) = \gamma[\delta k(\mathrm{t})^{-\rho} + (1-\delta)n(\mathrm{t}+s)^{-\rho}]^{-1/\rho}; \quad 0 < \delta < 1, -1 < \rho < \infty$$

where $\sigma = 1/(1+\rho)$. The marginal product of employment is

$$\partial y(\mathrm{t}+s)/\partial n(\mathrm{t}+s) = (1-\delta)\gamma[\delta(n(\mathrm{t}+s)/k(\mathrm{t}))^{\rho} + (1-\delta)]^{-(1+\rho)/\rho}.$$

The elasticity of the marginal product of employment with respect to the employment-capital ratio is

$$(11) \quad \partial \log(\partial y/\partial n)/\partial \log[n(\mathrm{t}+s)/k(\mathrm{t})] = -(1+\rho)\delta/[\delta + (1-\delta)(n(\mathrm{t}+s)/k(\mathrm{t}))^{-\rho}].$$

The concavity of the relationship between

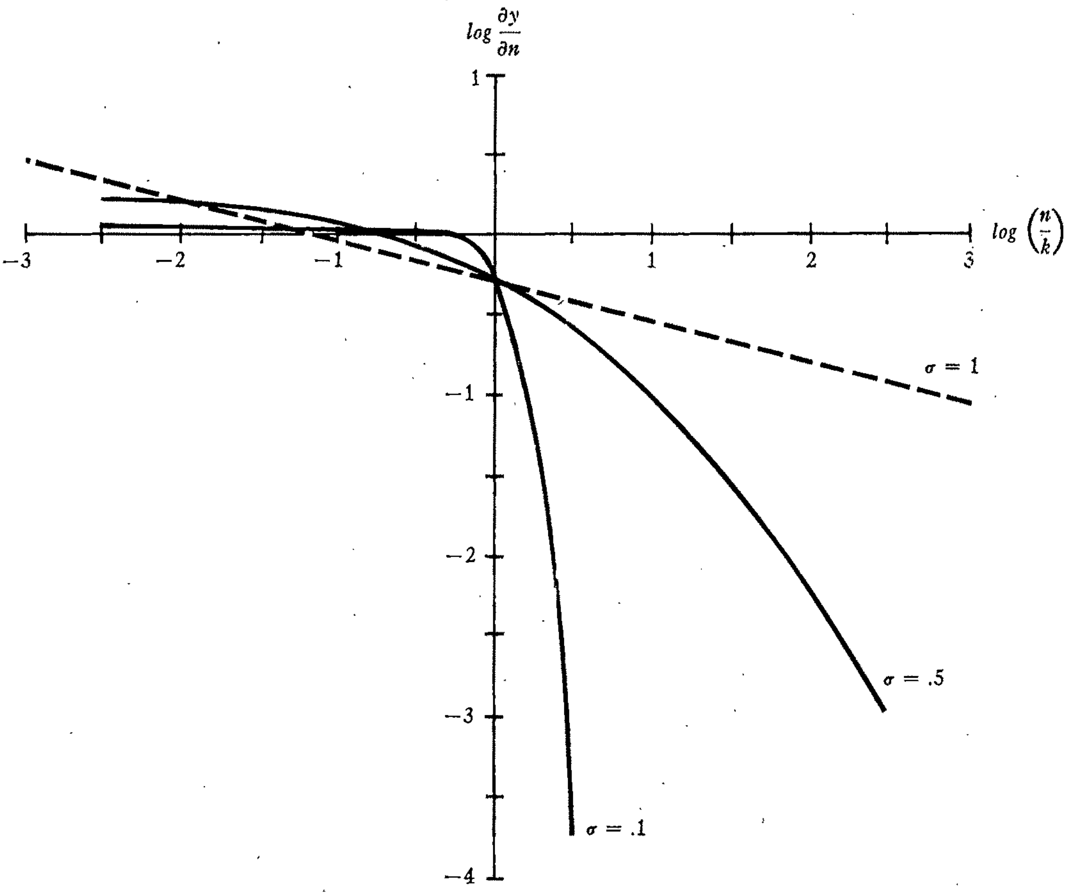


FIGURE 1 ($\delta = .25$)

log $(\partial y/\partial n)$ and *log* $[n(\mathrm{t}+s)/k(\mathrm{t})]$ is found by differentiating (11):

$$Sign\ \partial^2 \log (\partial y/\partial n)/\partial[\log (n(\mathrm{t}+s)/k(\mathrm{t}))]^2 = -Sign\ (1+\rho)\delta\rho(1-\delta)[n(\mathrm{t}+s)/k(\mathrm{t})]^{-\rho}$$

Thus, for a *CES* function with $\sigma < 1 (\rho > 0)$, the *log* of the marginal product of employment is a concave function of the *log* of the employment-capital ratio, as shown in Figure 1. For such a production function, a decrease in the straight-time real wage $w_0(\mathrm{t})$ will, with $\phi(s)$ constant, increase the ratio $n(\mathrm{t}+s)/n(\mathrm{t})$ for s for which $\phi(s) > 1$, and so increase the *relative* importance of high productivity workers. As it turns out, if the relationship between *log* $(\partial y/\partial n)$ and *log* (n/k) is sufficiently concave, if $w_0(\mathrm{t})$ lies in an appropriate range, and if $\phi(s)$ has an appropriate form, the derivatives (9) and (10) can both be negative, enabling Lucas's model to explain the empirical behavior of output per man-hour and average hourly real earnings. In addition, under the same conditions just listed, it is possible for the derivative of labor's share with respect to the straight-time real wage to be zero or negative, thus enabling the model to explain an acyclical or procyclical pattern for labor's share.[5]

---

[5] For a "one-shift" model, i.e., one in which $\phi(s) = 1$, for $0 \leq s \leq 1$, it is known that the elasticity of labor's share with respect to the real wage equals $1-\sigma$. See Murray Brown (pp. 185–6). Consequently, with σ between zero and unity, a one-shift model can't explain procyclical behavior of labor's share.

TABLE 1

σ	$w_0=0.7$			$w_0=0.8$			$w_0=0.9$		
	$\epsilon_{Y,MH}$	$\epsilon_{\bar{w},w}$	$\epsilon_{\xi,w}$	$\epsilon_{Y,MH}$	$\epsilon_{\bar{w},w}$	$\epsilon_{\xi,w}$	$\epsilon_{Y,MH}$	$\epsilon_{\bar{w},w}$	$\epsilon_{\xi,w}$
.05	.70	1.00	.95	.80	1.00	.95	.89	1.00	.95
.10	.70	1.00	.90	.79	1.00	.90	.88	1.00	.90
.15	1.40	−6.78	−5.84	.79	1.00	.85	.88	1.00	.85
.20	1.00	.05	.03	.79	1.00	.80	.87	1.00	.80
.25	.96	.41	.30	.79	1.00	.75	.86	1.00	.75
.30	.93	.56	.39	.78	1.00	.70	.85	1.00	.70
.35	.90	.64	.41	.78	1.00	.65	.84	1.00	.65
.40	.88	.70	.41	1.06	−1.13	−.77	.84	1.00	.60
.45	.86	.74	.40	.96	.25	.11	.83	1.00	.55
.50	.84	.78	.39	.92	.54	.25	.82	1.00	.50
.55	.82	.81	.36	.88	.67	.29	.94	.44	.17
.60	.81	.84	.33	.86	.76	.30	.90	.64	.24
.65	.80	.87	.30	.84	.81	.28	.87	.75	.25
.70	.79	.89	.26	.82	.86	.25	.85	.82	.24
.75	.78	.91	.23	.80	.89	.22	.83	.87	.21
.80	.77	.93	.18	.79	.92	.18	.81	.91	.18
.85	.76	.95	.14	.78	.94	.14	.79	.94	.14
.90	.76	.95	.10	.77	.97	.10	.78	.95	.10
.95	.75	.98	.05	.76	.98	.05	.76	.98	.05
1.00	.75	1.00	.00	.75	1.00	.00	.75	1.00	.00

We have calculated three pertinent elasticities for a two-shift version of Lucas's model where the instantaneous production function is the *CES* function

$$y(t+s) = [.25k(t)^{-\rho} + .75n(t+s)^{-\rho}]^{-1/\rho},$$

where $\sigma = 1/(1+\rho)$. We assumed that $\phi(s) = 1$ for $s < 0.5$ and $\phi(s) = 1.5$ for $s \geq 0.5$. For given values of the straight-time real wage $w_0(t)$ and σ, we calculated the elasticity of daily output with respect to man-hours, $\epsilon_{Y,MH}$; the elasticity of average hourly real earnings with respect to the straight-time real wage, $\epsilon_{\bar{w},w}$; and the elasticity of labor's distributive share with respect to the straight-time real wage, $\epsilon_{\xi,w}$. To explain the empirical observations for the postwar United States, $\epsilon_{Y,MH}$ must equal or exceed unity, $\epsilon_{\bar{w},w}$ must be zero or negative, and $\epsilon_{\xi,w}$ should be about zero.

Table 1 shows values of the three elasticities for values of σ between 0 and 1 and for straight-time real wages of .7, .8, and .9. For straight-time real wages of .7 and .8, there are values of σ for which $\epsilon_{Y,MH} > 1$, $\epsilon_{\bar{w},w} < 0$, and $\epsilon_{\xi,w} < 0$. Where one of these inequalities is satisfied, so are the other two. As the figures for a straight-time real wage of .9 show, however, there need not exist a σ for which these inequalities are satisfied at a given straight-time real wage. What is required for the three inequalities to hold is that the straight-time real wage lie in a region for which straight-time employment is subject to quite rapidly diminishing returns, with second-shift employment subject at the same time to much less rapidly diminishing returns.

Viewed as functions of σ given w_0, each of the three elasticities is characterized by a discontinuity. The discontinuity occurs at the value of σ at which 1.5 w_0 equals the marginal product of employment at zero employment, so that for this value of σ and those lower than it, employment only occurs on the first shift. For σ's lower than this critical value, then, the model behaves as does a standard one-shift model.

Table 1 shows that, given σ, each of the three elasticities depends on the straight-time real wage. For example, as the straight-time real wage changes, it is possible for output per man-hour sometimes to vary directly with man-hours, and sometimes to vary inversely with man-hours. Consequently, the model is in principle capable of explaining fairly complicated patterns in which, say, output per man-hour sometimes increases and sometimes decreases with manhours per unit of capital, depending on the range within which the straight-time real wage is moving.

REFERENCES

R. G. Bodkin, "Real Wages and Cyclical Variations in Employment: A Reconsideration," *Can. J. Econ.*, Aug. 1969, *2*, 353–374.

——— and L. R. Klein, "Nonlinear Estimation of Aggregate Production Functions, *Rev. Econ. Statist.*, Feb. 1967, *49*, 28–44.

M. Brown, *On the Theory and Measurement of Technological Change,* Cambridge 1966.

F. M. Fisher, "Aggregate Production Functions and the Explanation of Wages: A Simulation Experiment," *Rev. Econ. Statist.*, Nov. 1971, *53*, 305–325.

N. Georgescu-Roegen, "The Economics of Production," *Amer. Econ. Rev., Proc.*, May 1970, *60*, 1–9.

R. E. Hall and D. W. Jorgenson, "Tax Policy and Investment Behavior: Reply and Further Results," *Amer. Econ. Rev.*, June 1969, *59*, 388–400.

R. E. Lucas, Jr., "Capacity, Overtime, and Empirical Production Functions," *Amer. Econ. Rev., Proc.*, May 1970, *60*, 23–27.

J. Marchal and B. Ducros, *The Distribution of National Income,* New York 1968.

A Disequilibrium Model of Demand for Factors of Production

By SHERWIN ROSEN AND M. ISHAQ NADIRI*

Production technology and profit maximizing behavior imply that derived demands for factors of production are necessarily interdependent. Yet the literature on employment and investment functions has not fully appreciated or exploited this interdependence: most employment models assume fixed capital stock, while most investment models assume instantaneous adjustment of employment. In fact, factors of production are neither completely fixed nor perfectly variable, and time-adjustment patterns of different inputs are interconnected; thus both input stocks and input rates of utilization are interrelated. If all these interrelationships are properly considered, theoretical and empirical evaluation of input adjustment patterns, factor responses to price and output changes, and the distinctions between short- and long-run behavior all can be investigated in a mutually consistent manner.

In our research (see Nadiri and Rosen), we have developed a general model of factor adjustments which incorporates all potential interactions between stocks and utilization rates. Factors of production explicitly treated are employment of both production and nonproduction workers, hours worked per man, capital stock, inventories and a measure of capital utilization. The model is based on neoclassical production theory in the presence of adjustment costs. In this report we sketch the main features of the model, illustrate its usefulness in empirical estimation of factor demands, and present some empirical results.

I. The Model

The first problem considered is minimization of costs (1) subject to a production function (2):

$$(1)\quad C = y_1y_2w_p + y_1s_p + y_6y_7w_n + y_6s_n + y_3c + y_5c_I$$

$$(2)\quad x = A\prod_{i=1}^{7}(y_i)^{\alpha_i}, \qquad (i = 1 \ldots 7)$$

where y_1 denotes production workers, y_2 hours worked per man, y_3 capital stock, y_4 capital utilization, y_5 inventories, y_6 nonproduction labor, and y_7 hours worked by nonproduction employees. Output is denoted by x, and α_i are output elasticities of each input. Hourly wage rates for the two types of labor are given by w_p and w_n, while s_p and s_n are "rental" values associated with hiring such labor. Finally, c and c_I are, respectively, user costs of capital and inventories. No utilization rate is assumed for y_5, and the marginal effect of y_4 on c is positive. Constrained cost minimization yields a set of *log*-linear demand functions for stocks and utilization rates of inputs, written in matrix form as

$$(3)\qquad Y = k + \xi X + BR$$

In (3) all variables are in *logs*. Y is a vector of the *log* of desired input values, $X = \log x$, R is a vector of *logs* of factor prices, ξ is a vector of scale effects and B is a matrix of substitution effects. Since scale effects do not enter utilization demand functions, some elements of ξ may be zero.

---

* University of Rochester, and New York University and National Bureau of Economic Research, respectively.

Equation (3) represents long-run demand functions for factors. The dynamic adjustment paths of inputs toward desired target values are given (in *log*-linear terms) by a system of proportionate adjustment equations (4*i*):

$$(4i)\quad Y_{it} - Y_{it-1} = \sum_{j=1}^{7} \beta_{ij}(Y_{jt}^{*} - Y_{jt-1}) + \epsilon_{it}; \qquad (i = 1, \ldots, 7)$$

where the β_{ij}'s are adjustment coefficients, the ϵ's are stochastic terms, and $Y_{it} = log\ y_{it}$. An important feature of (4*i*) is that the matrix $[\beta_{ij}]$ is *not* specified in advance to be diagonal. Hence there is the possibility of feedbacks and interrelated adjustments among all factors of production. The theoretical motivation of specification (4*i*) derives from rational firm behavior in the presence of adjustment costs. Moreover, since inventories are endogenous to this system, sales instead of output is used in equation (3). Distinctions between transitory and permanent components of demand are also introduced.

Production function constraints to the effect that firms are on their production frontiers imply some restrictions on adjustment coefficients β_{ij}. In particular, the matrix $[I-\beta]$ should be singular, because all adjustments cannot be independent. Given sales and the production function, any six decisions among the inputs implies the seventh. In addition, dynamic adjustment properties of each input depend on the nature of this matrix. The configuration of elements in $[I-\beta]$ determines factor adjustment patterns and the results reported in the literature all are special cases of this model. Particularly, the generalized adjustment process embedded in (4*i*) allows for adjustment responses that are inadmissible under familiar independent stock adjustment specifications. In particular, if firms are out of long-run equilibrium with respect to any one input, a corresponding disequilibrium must appear elsewhere in the system, for otherwise the sales constraint could not be maintained. Therefore, some inputs temporarily can overshoot their long-run values in order to maintain sales targets when other inputs are more difficult and costly to adjust. These possibilities are unique to our model and cannot be captured by alternative specifications.

II. The Estimates

Quarterly time-series data for total durable and nondurable goods industries used to estimate the model come mostly from published sources, but time-series observations for capital stock, capital rental values, and utilization rates had to be constructed. The variables s_n and s_p have been omitted due to lack of adequate data. Also, wages of nonproduction workers were not available and are assumed to vary in proportion to wages of production workers. We used the sample period 1948I to 1967IV to estimate the model. (We obtained results similar to those given below when we used disaggregated industry data; see Nadiri and Rosen.)

Substitution of (3) into (4*i*) yields the structural equations to be estimated:

$$(5i)\quad Y_{it} = a_{i0} + a_{i1}S_t + a_{i2}\left(\frac{W}{C}\right) + a_{i3}T + b_{i1}Y_{1t-1} + b_{i2}Y_{2t-1} + \ldots + b_{i6}Y_{6t-1} + U_{it}, \qquad (i = 1, \ldots, 6)$$

Sales, S, relative factor prices (W/C) and trend, T, are exogenous; a_{ij} and b_{ij} are regression parameters; t subscripts refer to time. Serial correlation of estimated residuals U_{it} has been taken into account by Cochrane-Orcutt transformations of each equation. Generalized least squares estimates of the structure (5*i*) for total durables and nondurables are shown in Tables 1 and 2. Each column of these tables gives an estimate of the corresponding equation in (5*i*). The coefficients of sales, trend and relative prices indicate one quarter response elasticities while coefficients on lagged dependent variables estimate adjustment terms in the matrix $[I-\beta]$ in equation (4*i*).

Note that the goodness of fit of each equation is very high. Indeed, our model performs significantly better than a set of naive autoregressive models and post-sample conditional forecasts of the model proved (in results not reported here) superior to alternative autoregressive specifi-

TABLE 1—ESTIMATED STRUCTURE OF MODEL (5i), TOTAL DURABLES
Sample period: 1948I–1967IV

Independent Variables	Dependent Variables					
	Production Labor Y_{1t}	Hours Y_{2t}	Capital Y_{3t}	Utilization Y_{4t}	Inventories Y_{5t}	Nonproduction Labor Y_{6t}
Constant	−2.911	1.165	.2978	2.983	−1.050	−.9400
	(2.582)	(2.703)	(1.285)	(1.126)	(.7102)	(1.373)
Sales (*S*)	.4094	.1271	.0015	1.069	.0829	.0345
	(12.49)	(9.788)	(.2667)	(13.36)	(2.144)	(2.059)
Trend (*T*)	−.0041	−.0001	.0005	−.0067	.00005	.0012
	(5.337)	(.5841)	(1.453)	(3.896)	(.0398)	(1.251)
Relative Prices (*W/C*)	−.0597	−.0137	.0013	−.2046	−.0085	−.0355
	(2.130)	(1.341)	(.1993)	(3.265)	(.1970)	(1.742)
Y_{1t-1}	.5472	−.0628	.0477	−.4448	.2610	−.0436
	(8.236)	(2.570)	(3.076)	(2.963)	(2.751)	(.9545)
Y_{2t-1}	.6860	.7016	−.0216	−.7688	.3957	.3090
	(2.566)	(6.840)	(.4503)	(1.219)	(1.156)	(2.166)
Y_{3t-1}	.0921	−.0733	.8822	−.3538	−.0240	−.0480
	(1.039)	(2.241)	(24.46)	(1.761)	(.1633)	(.4543)
Y_{4t-1}	−.1186	−.0565	−.0059	.3560	−.0508	.0208
	(4.343)	(5.568)	(1.147)	(5.711)	(1.382)	(1.347)
Y_{5t-1}	−.1180	−.0394	.0186	−.5077	.5980	.0373
	(2.415)	(2.223)	(1.408)	(4.661)	(7.989)	(.9578)
Y_{6t-1}	.1374	.0301	.0444	.6656	.3040	.8126
	(2.693)	(1.668)	(1.594)	(5.999)	(3.197)	(9.972)
R^2	.9827	.9589	.9999	.9199	.9972	.9994
$\hat{\rho}$	.1721	.0003	.9059	−.0069	.5925	.9024
SEE	.0118	.0048	.0020	.0300	.0137	.0059
SSR	.0097	.0016	.0002	.0622	.0131	.0024

Note: All variables in logs except trends. t-statistics are in parentheses. R^2 is coefficient of determination. *SEE* is standard error of estimate and *SSR* is sum of squared residuals. $\hat{\rho}$ is estimated serial correlation coefficient.

TABLE 2—ESTIMATED STRUCTURE OF MODEL (5i), TOTAL NONDURABLES
Sample period: 1948I–1967IV

Independent Variables	Dependent Variables					
	Production Labor Y_{1t}	Hours Y_{2t}	Capital Y_{3t}	Utilization Y_{4t}	Inventories Y_{5t}	Nonproduction Labor Y_{6t}
Constant	−2.870 (4.894)	.8273 (1.817)	.2316 (1.145)	−2.599 (1.461)	1.109 (1.135)	−1.352 (5.089)
Sales (S)	.2657 (4.956)	.2126 (4.510)	−.0095 (.6645)	.9106 (4.547)	−.0537 (.5470)	−.0197 (.6437)
Trend (T)	−.0024 (3.699)	−.0016 (3.391)	.0001 (.4823)	−.0092 (4.593)	.0021 (1.971)	.0001 (.5084)
Relative Prices (W/C)	−.0004 (.0273)	.0115 (.8776)	.00006 (.0106)	.0381 (.7786)	−.0050 (.1756)	−.0060 (.8415)
Y_{1t-1}	.5935 (9.218)	−.0866 (1.925)	.0582 (2.188)	−.2063 (1.230)	.3971 (4.023)	−.0261 (1.056)
Y_{2t-1}	.5008 (5.451)	.6033 (4.873)	−.0288 (.7241)	.3704 (.7330)	.0766 (.2938)	.3679 (4.815)
Y_{3t-1}	.2583 (3.540)	.0316 (.6764)	.9371 (26.65)	−.1982 (1.156)	−.2940 (2.824)	.0359 (1.419)
Y_{4t-1}	.0098 (.4254)	−.0353 (1.602)	−.0035 (.6141)	.1391 (1.389)	−.0755 (1.682)	−.0007 (.0447)
Y_{5t-1}	−.0109 (.2723)	−.0299 (.9989)	.0058 (.4519)	−.4459 (3.807)	.6518 (10.04)	.0088 (.5014)
Y_{6t-1}	−.2268 (2.654)	−.0058 (.0951)	.0645 (1.480)	1.145 (4.635)	.4601 (3.465)	.9385 (24.95)
R^2	.9634	.8618	.9998	.7928	.9964	.9992
$\hat{\rho}$	.5708	.2780	.8878	.0234	.3532	−.0727
SEE	.0062	.0056	.0017	.0255	.0115	.0040
SSR	.0026	.0021	.0002	.0450	.0092	.0011

Note: All variables in *logs* except trend. t-statistics are in parentheses. R^2 is coefficient of determination. *SEE* is standard error of estimate and *SSR* is sum of squared residuals. $\hat{\rho}$ is estimated serial correlation coefficient.

cations for 1968I to 1969IV.

The short-run impact elasticity estimates generally are statistically significant. The immediate response of utilization rates to sales is very high compared to that of other inputs. Production employment and hours also exhibit strong sales impacts but of lesser magnitudes. On the other hand, initial impacts of relative prices are generally in the right direction but are very small and often statistically insignificant. The full dynamic adjustment specification (4*i*) is borne out by the data. Own-adjustment coefficients $\beta_{ij}(i=j)$ —the estimated regression coefficients of the lagged endogenous variables on the diagonal of the tables—are all positive, statistically significant and less than unity;

cross-adjustment coefficients $\beta_{ij}(i \neq j)$—off-diagonal regression coefficients of lagged endogenous variables—are usually statistically significant, especially in total durables, an important verification of the "feedback" specification (4*i*). Generally speaking, production labor interacts with hours, capital stock, utilization and inventories; nonproduction labor interacts with production worker utilization and inventories; and capital stock has strong feedback with some labor variables. As a rule, the influence of disequilibrium in stocks of inputs on utilization rate variables is much more powerful than the reverse, implying that utilization rates act as buffers in maintaining sales targets, while stock inputs slowly adjust.

These patterns of adjustments indicate that utilization rates are truly variable inputs and that there is a hierarchy of adjustment speeds among stocks, ranging from fairly rapid adjustment of production employment to rather slow adjustments of capital stock. The distributed lag responses implied by the coefficients in Tables 1 and 2 overwhelmingly verify the specification. Space limitations preclude exhibiting them here but the main features with respect to changes in sales may be briefly outlined: (a) Hours of work and the utilization rate overshoot their long-run values almost immediately after an increase in sales, again indicating their buffer role in the input adjustment process. (b) Production labor exhibits a general exponential adjustment pattern, with most of the response taking place three to four quarters after sales change. (c) Distributed lag adjustment patterns of inventories and nonproduction workers display some tendency to overshoot their ultimate long-run values and the response usually builds up for two or three quarters after changes in sales occur. (d) Capital stock is the most slowly adjusting input and exhibits the familiar bell-shaped response pattern. (e) Finally, the overall adjustment response pattern in durables is more rapid than the corresponding pattern in the nondurables.

Long-run responses to changes in the exogenous variables computed from the estimates in Tables 1 and 2 are shown in Table 3. It is clear that long-run elasticities of demand differ widely between the two industries. However, they have some major features in common. In both sectors long-run price elasticities—the substitution terms in (3)—are very small both absolutely and relative to sales responses. Evidently, short-run scale effects dominate

TABLE 3—LONG-RUN ELASTICITIES
(Computed from Tables 1 and 2)

	Production Labor Y_1	Hours Y_2	Capital Y_3	Utilization Y_4	Inventories Y_5	Nonproduction Labor Y_6
Durables						
Price	.019	.042	−.035	−.522	−.032	−.168
Trend	−.004	.001	.006	−.009	.004	.007
Sales	.701	.002	.411	.544	.646	.101
Nondurables						
Price	.046	.018	.059	.054	.013	.024
Trend	−.016	−.001	−.022	−.013	−.003	−.010
Sales	2.063	.118	3.474	1.436	1.219	1.694

in our data, rendering substitution effects very difficult to estimate. Effects of trend are consistently negative in nondurables and of mixed signs in durables. This variable is a conventional proxy for technical change and unmeasured factor improvements over the sample period, but because there was a marked trend in the relative wage-rental ratio (W/C) over the period it may also capture part of the price effect. Finally, there seems to be evidence of increasing returns to scale in durables and of decreasing returns to scale in nondurables. In interpreting this last result, note that we have not artificially imposed all the restrictions of our model on the data (viz. $|I-\beta|=0$) and the production function parameters are overidentified. As a check on this restriction we computed the characteristics roots of $(I-\beta)$ and in both sectors, especially nondurables, the smallest root was reasonably close to zero (.16 for durables and .05 for nondurables). Insofar as the restrictions do not hold identically, particular point estimates in Table 3 must be interpreted with caution.

Given the appropriate qualifications on identification, adjustment response differences between the two sectors may be due to the role of inventories. In nondurables, finished goods inventories are relatively more important and serve as better buffers against sales fluctuations than do input inventories. In durables, however, finished goods are less homogeneous because of the technical specificity of final output in that sector. Such differences in the nature of output link sales and input decisions more closely in durables. Therefore, adjustment speeds are more rapid in the durables. In addition, Table 3 suggests a smaller degree of returns to scale in nondurables. Hence, costs rise more rapidly in that sector, making production smoothing over the business cycle more economical.

III. Applications

System ($5i$) represents a generalized model of factor demand in time-series and includes all existing formulations as special cases. For example, conceptually we may "fix" capital stock, utilization, inventories, and nonproduction workers in equations ($5i$) and then generate short-run demands for man-hours and calculate short-run labor productivity. Repeating the experiment and sequentially relaxing each constraint by allowing some other factors to "vary," we can generate alternative measures of labor productivity. Our empirical results indicate that others have obtained such large estimates of returns to labor probably because their short-run employment functions do not capture effects of variation in capital utilization. Thus, their estimates of short-run labor productivity should not really be regarded as returns to labor alone.

In our model the price and scale effects are separately estimated and can be different for the stock and utilization rate variables. The results indicate that substitution effects are dominated by scale effects not only at the aggregate level presented above but also at the two-digit disaggregated level. The model has no restrictions which would automatically produce such results; thus, depending on the quality of our time-series data, our estimates of price elasticities are either biased or truly close to zero.

Finally, the structure of adjustment embedded in our model is very flexible; it allows for overshooting, indicating the interdependence of input decisions. These overshoots and feedbacks are extremely important for short-run stabilization policies. Should policymakers concentrate their attention on only one of the decision variables, say investment, they may ignore the compensatory effects of other variables, such as employment and utilization, that vary sympathetically to maintain output. These effects are of substantial magnitude; and results obtained by fitting independent input equations are conceptually inadequate and can lead to serious policy errors.

REFERENCE

M. I. Nadiri and S. Rosen, *A Disequilibrium Model of Demand for Factors of Production,* New York, forthcoming, 1974.

THE POLITICAL ECONOMY OF MULTINATIONAL ENTERPRISES

Multinational Corporations and the International Adjustment Process

By G. C. HUFBAUER*

The close of the Second World War ushered in an era of freer trade, fewer restrictions on capital movements, and the more rapid diffusion of technology among the industrial nations. The multinational corporation was an important, but not the only, actor in this drama. Under the auspices of the International Monetary Fund and the General Agreement on Tariffs and Trade, governments have negotiated convertibility and lowered trade barriers, and through cooperation in scientific and technical fields, government agencies and universities have hastened the diffusion of knowledge. How do these institutional changes modify the international adjustment process? I shall first consider the long-run integration of national economies and the relocation of industry from one geographic area to another. I shall then turn to the short-run question of balance of payments disequilibria.

I. Adjustment in the Long Run

Restrictions on the flow of technology and capital have come down more quickly than barriers on trade. Multinational enterprise is both a cause and consequence of this sequence. Numerous surveys and statistical studies (summarized in Hufbauer) suggest that trade barriers have prompted the spread of multinational corporations. Manufacturing firms take their capital and knowledge abroad to produce locally for the protected market, rather than export home goods or forego sales. Thus, in the fashion of Robert Mundell, the export of capital and technology has substituted for the export of products—defensive investment, it is called. In the bargain, production technology has begun to resemble its idealized neoclassical portrait: a beneficent climate equally available to all countries.

In contrast with their defensive investment patterns of the past, multinational corporations might next embark on a wave of investment rationalization, at least within the industrial countries. This prognosis critically depends both on the continued relaxation of trade barriers and on the continued availability of cheap transport. The relaxation of trade restraints can hardly be expected in an era of worldwide inflation which brings an unwelcome round of export controls. Nor can cheap transport be assured in the midst of an energy crisis. Without a lessening of the frictions which impede trade, there can be no concentration of production in locations of least cost. Yet, despite the unsettled world scene, there are signs of rationalization. Between 1966 and 1970, the share of manufacturing output exported by all foreign affiliates of American multinational firms rose from 26 percent to 30 percent (U.S. Department of Commerce). In the same period, *U.S.* imports from less-developed countries under sections 806.30 and 807.00 of the U.S. Tariff Schedules, which permit import duties to be levied only upon value added abroad when inputs originate in the United States, rose from $61 million to $539 million (Gerald K. Helleiner). Much of this trade was sponsored by multinational firms.

---

* Professor of economics, University of New Mexico, on academic leave to King's College, Cambridge, with the assistance of a Fulbright-Hays award. I am grateful to R. Pearce's seminar at Reading University, to D. McAleese's seminar at Trinity College, Dublin, and to M. R. Killingsworth for critical comments on earlier drafts.

Suppose that trade liberalization continues, unimpeded by protective impulses, export controls, or an energy crisis. What effect will the ensuing rationalization exert on the degree of specialization and the volume of trade? What will these changes mean for labor and its role in the long-run adjustment process?

Elsewhere, Hufbauer and J. G. Chilas have shown that specialization between the four regions of the United States in the production of manufactured goods is far more extensive than specialization between the major industrial nations or even within the European Common Market. It thus seems reasonable to suppose that the industrial nations will become very much more specialized than they are today. Multinational firms could well lead the rationalization process. Once trade barriers are removed, the multinational enterprise has the easy alternative of supplying a market from its own factory abroad. Unlike competitors operating from a single country, it need not engage in ruinous price wars to determine which plant is most efficient in which product. As an example of things to come, the giant Swedish firm, SKF, manufactures specialty ball bearings in Sweden, where the labor force is highly skilled, and makes longer runs of standard ball bearings in Britain and Germany, principally for the automotive industry.[1]

In a world of multinational corporations, with no restriction on the flow of capital, technology, or goods, and a growing degree of specialization, the relative immobility of labor could lead to serious adjustment problems. Again, the comparison between interregional and international events is instructive. Between 1960 and 1970, rates of net migration between regions were substantially higher than rates between independent Western countries of approximately the same size.[2] Such comparisons suggest that prevailing rates of international migration are not large enough to relieve the potential strains created in the labor market by a combined regime of free trade in goods, capital, and technology. The prevailing rates are not only too small for the sort of continuous adjustment that regularly occurs within a national economy; they are even less adequate for the quantum leap in specialization which might accompany a regime of free trade.

In a setting of limited labor mobility, the market mechanism relieves imbalances in labor surplus regions through declining wage/profit and wage/rent ratios, leading to greater emphasis on labor-intensive production techniques, and, later, on labor-intensive goods. The opposite happens in regions of labor scarcity. But these adjustments can be unpleasant. Some areas must experience a wage decline relative to others. Moreover, as the labor-surplus region adopts labor-intensive techniques and specializes in labor-intensive goods, the outward flow of capital to regions of labor scarcity will be accelerated (or the inward flow reduced). Inadequate labor mobility between regions may thus magnify the requisite flows of capital. It is hard to imagine the government of a depressed, labor-surplus region tolerating an increased outflow of capital without compensation.

---

[1] K. Kojima contends that Japanese direct investment is already trade oriented. Japanese firms concentrate on the exploitation of raw materials and on the overseas manufacture of labor-intensive goods, and much of the output is then shipped back to Japan. By contrast, American firms often manufacture their most sophisticated products for locally protected markets. If Kojima's description is correct, rationalization of multinational investment could come to resemble the Japanese pattern.

[2] A comparison of net migration experience between 1960 and 1970 shows that a collection of Western countries had decadal net migration rates averaging between 1.9 percent and 2.3 percent of the base year population, whereas regions of approximately the same size had rates averaging between 3.4 percent and 3.9 percent. (The figures are available on request.)

These unpleasant features virtually compel government adjustment assistance on an international basis. The alternatives are faster international migration or slower liberalization of trade. Adjustment programs will have to be launched on a grand scale if they are to succeed. The European Economic Community has tentatively announced a program which would transfer some £330 million annually to needy regions, defined to encompass one-third the population of Europe (*The Times*). This is not the place to embark on detailed calculations. But by comparison, the *U.K.* regional program has an annual budget cost of at least £300 million. It seems unlikely that the adjustment problems of Europe can be solved on a thinly spread *U.K.*-sized budget. And discussions have yet to begin on a multilateral adjustment program which matches the farflung scope of multinational corporations.

Apart from the size of adjustment assistance, the method of administration is important. As operated within the industrial countries, regional adjustment programs usually subsidize the employment of both capital and labor in labor-surplus areas (B. Moore and J. Rhodes; Commission of the European Communities, 1971, 1973). But if specialization creates unemployment, labor subsidies alone would seem appropriate. Capital subsidies may prove counterproductive by delaying the transition to labor-intensive products and processes in labor-surplus areas.

II. Adjustment in the Short Run

Suppose that ambitious and sensible adjustment programs are launched and that economic integration coupled with the spread of multinational enterprise proceeds apace. What will be the nature of short-run adjustment mechanisms in the new era?

Presently there are three major kinds of policies for restoring short-run payments equilibrium. The familiar elasticities approach focuses on relative commodity prices, both between domestic and foreign products and between traded and nontraded goods. To correct a deficit, policy measures must shift relative prices enough so that the payments gap is closed by an improvement in the trade balance.

The absorption approach emphasizes aggregate savings. According to the Keynesian algebra of national accounts, a balance of payments deficit implies negative foreign investment and inadequate domestic savings. Policy must seek to enlarge domestic savings at the expense of domestic absorption of output.

Finally, the monetary approach (synthesized by Harry G. Johnson) stresses the relationship between the demand for money, the stocks of real and financial assets, the flow of real output, and the price level. The monetary approach assumes high mobility of capital between nations: thus a single interest rate will rule in the international capital market. It also assumes that *relative* commodity prices are little affected by currency appreciation or depreciation even though the *level* of prices may change. Two kinds of money are available to the nation, domestic and for-

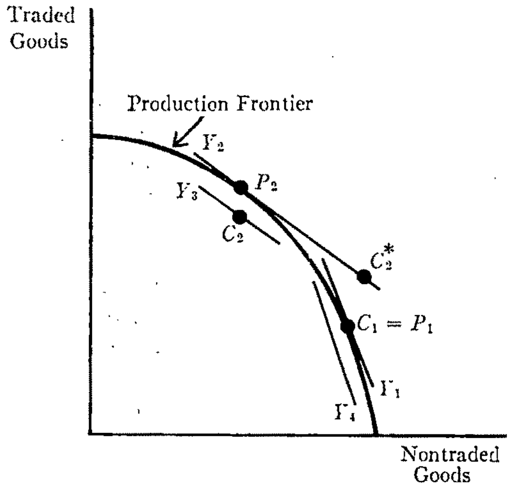


FIGURE 1

eign, and each kind can be converted to the other. To correct a deficit, policy measures must stimulate home demand for foreign money. This can be achieved by reducing the nominal supply of domestic money, by raising the domestic price level, or by increasing the level of output.

Figure 1 illustrates the key ideas. Full employment is assumed throughout. The initial real expenditures line and price ratio is given by Y_1, and consumption and production occur at point $C_1 = P_1$. Since production equals consumption of traded goods, the balance of trade is zero. Suppose, however, that the balance of payments is in deficit because of a capital outflow. The country devalues, thereby switching production to P_2 and *attempted* consumption to C_2^* (both traded and nontraded goods have a positive income elasticity). Attempted consumption leaves a surplus of traded goods; and according to the elasticities approach, this surplus can be used to improve the balance of trade. But the elasticities approach neglects the deficit in attempted consumption of nontraded goods.

At this juncture, the absorption school comes on stage: fiscal policy must be invoked to depress the real expenditures line to Y_3, where consumption of nontraded goods at C_2 equals production at P_2 (meanwhile, there is still a surplus of traded goods). Here the monetary protagonists enter the drama. Devaluation itself, they say, depresses the expenditures line. Devaluation causes prices to rise, firstly among traded goods and secondly, if excess demand still exists, among nontraded goods. The higher price level reduces real expenditures as the private sector attempts to restore the real value of its monetary holdings. Goods are sold for foreign exchange, which is then converted to domestic money. Devaluation works partly through expenditure switching, but largely through domestic inflation and thence expenditure reduction. In the extreme case, all goods are traded, and devaluation works in much the same way as domestic monetary contraction: the expenditures line is reduced to Y_4 in Figure 1 with no change in its slope.

The integration of national economies compels greater emphasis on a monetary approach to short-run adjustment problems. Private capital flows have already become important components of the external balance among industrial nations. Adjustment can no longer be analyzed solely in terms of the trade account; yet, unlike the monetary approach, the elasticities and absorption approaches are disturbingly oblivious to the capital account. Moreover, with a vast exchange of intermediate goods, the exports of one area become costs of production in another. The ratio between import and export prices becomes increasingly resistant to change through currency realignment. Meanwhile, the nontraded goods sector shrinks: tourism grows, real transport costs decline and multinational enterprises bid up cheap electricity, office buildings and labor wherever they are available. Exchange-rate policy gradually loses its power to alter relative commodity prices. Thus, the elasticities approach begins to fade. To be sure, relative prices will still change—but in response to underlying productivity trends, shifts in demand and alterations in the monopolistic organization of markets rather than to exchange-rate variation. If governments wish to alter relative prices, they will increasingly have to rely on the imposition of selective taxes, subsidies, and controls. Selective restrictions are the last resort of nationalism, and, however politically popular, they cannot be reconciled with the ultimate integration of markets.

In the new world of integrated commodity markets and unified capital markets, the *overall* external balance may be determined solely by the difference between the growth of domestic money income and the growth of domestic credit. However, the *composition* of the external balance, as between the trade account and the capital account, will still be very much affected by fiscal and exchange-rate measures. A government deficit, financed by the issue of bonds, will worsen the trade balance and improve the capital balance. Correspondingly, devaluation will improve the trade balance and worsen the capital balance, although the impact may be attenuated by the effect of devaluation on investment. Further, until commodity markets become perfectly integrated, and the distinction between traded and nontraded goods disappears, exchange-rate policy can be usefully deployed to offset wage increases which make home goods less competitive on world markets. (These ideas are given in the clothes of a simple model outlined in an appendix available from the author on request.) Thus, so long as the composition of the external account matters and so long as national commodity markets are imperfectly integrated, fiscal and exchange-rate policies will remain important balance of payments instruments.

REFERENCES

G. K. Helleiner, "Manufactured Exports from Less-Developed Countries and Multinational Firms," *Econ. J.*, Mar. 1973, *83*, 21–47.

G. C. Hufbauer, "The Multinational Corporation and Direct Investment," Int. Finance Sect., Princeton Univ., Conference on Res. in Int. Trade and Finance, Mar. 1973.

——— **and J. G. Chilas**, "Specialization by Industrial Countries: Extent and Consequences," *Weltwirtschaftliches Archiv*, forthcoming.

H. G. Johnson, "The Monetary Approach to Balance-of-Payments Theory," *Intermountain Econ. Rev.*, Fall 1972, *2*, 1–13.

K. Kojima, "A Macroeconomic Approach to Foreign Direct Investment," *Hitotsubashi J. of Econ.*, June 1973, *14*, 1–21.

B. Moore and J. Rhodes, "Effects of British Regional Economic Policy," *Econ. J.*, Mar. 1973, *83*, 87–110.

R. A. Mundell, "International Trade and Factor Mobility," *Amer. Econ. Rev.*, June 1957, *47*, 321–335.

Commission of the European Communities, *Regional Development in the Community: Analytical Survey*, Luxembourg 1971.

———, *Report of the Regional Problems in the Enlarged Community*, Brussels 1973.

The Times, Oct. 13, 1973.

U.S. Department of Commerce, *Special Survey of U.S. Multinational Companies*, Springfield, Virginia, 1972.

Competition Policy Toward Multinational Corporations

By Raymond Vernon*

Multinational enterprises appear mostly in industries dominated by comparatively large firms and characterized by high entry barriers (see F. M. Scherer, John H. Dunning, and Vernon 1971). The problems of competition policy associated with multinational enterprises, therefore, are mainly the familiar ones associated with oligopoly behavior. Two features, however, are distinctive. First, the oligopoly's leadership generally includes firms operating from quite different national bases. Second, the interaction of the leaders typically spreads over many national jurisdictions.

I. Oligopolies Distinguished

Though multinational enterprises are largely formed in oligopolistic industries, the entry barriers associated with industries in which multinational enterprises predominate are of several different kinds.

In some industries, the barrier stems from the sheer size of an efficient unit of production. In principle, the necessary technology, capital, and organization required for such units may be available at a price, but in practice they may not be. Industries operating under these conditions—such as aluminum smelting and international crude oil—are generally long established and their products well standardized. These are the mature oligopolies. In a second type of oligopoly—underemphasized in the literature—entry requires the accumulation of information and experience that cannot be bought, and aspirants unavoidably face a long, uncertain period of learning (D. L. McQueen, Vernon 1970). The barrier commonly is present, for example, in advanced chemicals, complex electronics, and new machinery. Since such industries are usually associated with high expenditure on research-related activities, they are often dubbed technology-intensive or science-based industries. (A third type, sometimes difficult to distinguish and omitted here for lack of space, is built on the monopoly of trade names and advertising.)

II. The Multinationalizing of Mature Oligopolies

The leaders in mature oligopolies are rarely free of incipient threats to their position. A technological development may render their product obsolete (consider nylon's threat to rayon); a new process, such as the extraction of sulphur from crude oil, or the appearance of new entrants, eager to participate in the oligopoly rent by capturing a share of the market, may also be dangerous. Finally, one of the leaders may see an advantage in turning briefly from cooperative to competitive tactics.

Before World War II, leaders in the oligopolistic industries commonly tried to promote cooperative behavior by forming an industry cartel and dividing the existing markets among themselves. If the leaders were located in different countries, the market division was usually on geographical lines (Ervin Hexner). As a rule, these efforts managed to contribute a little to stability. In primary metals, for instance, I can find only one major *U.S.* firm that established a new manufacturing

---

* Harvard University.

beachhead in Europe during the period between the two great wars (J. W. Vaupel and J. P. Curhan 1969, p. 28). But in many industries, uncommitted newcomers often upset the arrangements. For instance, even though cartel agreements blanketed practically every conceivable chemical product during the interwar period, nine leading *U.S.* chemical firms in those years established their first manufacturing facilities in Europe (Vaupel and Curhan 1969, p. 22).

Prewar stability was generally based on cartels and articulated largely in market divisions—based on geographic lines (Hexner). However, it was undermined during the decade or two following World War II. Some cartels did reappear (Corwin D. Edwards), but they seemed a good deal less pervasive and less powerful. In general, the trade and investment links between markets in the advanced countries greatly increased (see Organization for Economic Cooperation and Development (*OECD*) 1972, U.S. Dept. of Commerce 1960, 1966, 1972). Europeans began anxiously to comment on the fact that substantial portions of their key local industries were being managed as part of a global system conducted by Americans.

The net effect of this tendency was to place a considerable proportion of *U.S.* manufacturing production itself in enterprises that were part of a global network. In various mature industries—including motor vehicles, fabricated metals, petroleum refining and rubber—over 50 percent of the sales in the *U.S.* market came to be accounted for by *U.S.*-based enterprises with multinational networks (Vernon 1971, p. 14).

At the same time, European firms in industries of this sort were beginning to show some interest in counterpenetration. At the end of 1971, in roughly the same group of industries, large multinational enterprises headquartered outside the United States controlled about 125 manufacturing subsidiaries in the United States, most of them being of very recent vintage (Vaupel and Curhan 1973). Japanese firms had not yet begun to participate in patterns quite like those of the Americans and the Europeans, but various signs already existed that they were approaching that stage (Yoshi Tsurumi).

III. Cooperation and Competition in the Mature Oligopolies

Since World War II, these tendencies have turned many national oligopolies into multinational oligopolies, especially in the mature industries. In these multinational oligopolies, as in those of national scope, the leaders have shared a common capacity for mutual destruction, a common concern over new entrants, and a common need to find some form of cooperative behavior. The one important difference is that the leaders no longer could protect themselves by agreeing to divide up markets on a simple geographical basis. As a partial alternative, they sought to join one another in joint ventures, follow one another into new locations, and establish bridgeheads in one another's main markets.

Potential competitors who participate in such joint ventures often want to capture the available engineering economies of production scale without having to make too large a commitment in any single foreign country; this is commonly the case for investments in copper mines, aluminum smelters, chemical plants, and automobile plants. But partnerships are also a means by which leaders can move toward a common set of cost structures and can observe one another's competitive behavior at closer range. And the incidence of such arrangements appears on the increase (see Vaupel and Curran 1973 and European Economic Community report).

Follow-the-leader behavior is more subtle and more difficult to measure. It consists essentially of the leaders matching one another's moves in establishing production facilities in new locations, to a degree that is not explained simply by their independent recognition of economic opportunities. The phenomenon so far is documented only with regard to the behavior of *U.S.*-based multinational enterprises (F. T. Knickerbocker).

The use of countervailing bridgeheads, sometimes called the exchange-of-hostage syndrome, may be less common than the follow-the-leader pattern, but there is not much doubt as to its existence. Foreign-owned enterprise A establishes a producing facility in B's home territory; B, fearing that A may be disruptive in its competitive behavior in that territory, establishes a producing facility in A's home territory. The object is partly to share A's strengths and partly to encourage A's cooperation. According to preliminary indications from my own research, this kind of behavior also seems on the increase.

Such growing internationalization of oligopoly in some industries might lead one to suppose that international markets in those industries are growing less competitive. The evidence, however, runs the other way. In recent decades, for instance, the mature oligopolies have been growing less concentrated in industrial structure, not more so. Though economies of scale in an absolute sense are probably increasing in most of these industries, other developments seem to be more than offsetting: both world markets and the number of enterprises capable of overleaping the barriers to entry are growing.

Some of the effects of these developments upon the structure of the industries are traced in Figure 1. That figure presents a measure—the well-known Herfindahl index—reflecting the degree of concentration of production among firms in four mature oligopolies. The index covers production outside the communist countries and other countries with state-controlled production facilities; in each case, the term "firm" comprises both the parent and its subsidiaries, wherever located. Similar measurements for other industries are in process, and it appears likely they will generate parallel results.

The decline in concentration displayed in Figure 1 does not inevitably suggest an increase of competition in world markets. In many countries, especially the European Economic Community (*EEC*) members, mergers and amalgamations have tended to increase concentration in many industries (*EEC*, p. 141). That tendency may be both a cause and a consequence of the trend portrayed in Figure 1: a cause in the sense that such mergers were the means by which medium-sized firms came to challenge the leaders; a consequence in the sense that increased competition may have shaken out many smaller firms previously sheltered in national markets.

In any event, once a new equilibrium has been achieved in world markets, the new leadership may have just as much incentive as the veterans to reestablish some semblance of cooperative behavior. Accordingly, after the period of expansion and growth, stabilizing techniques such as the three described earlier could become dominant. The increase in numbers may make agreement somewhat more difficult to achieve than in the past, but it cannot be assumed that the problem of numbers will defeat the urge among the leaders to cooperate.

IV. The Innovation-based Oligopolies

In the fields of pharmaceuticals, advanced chemicals, complex electronics, and intricate machinery, the imperfect nature of international competition stems mainly from the ability of producers to differentiate their products, fortified at times by

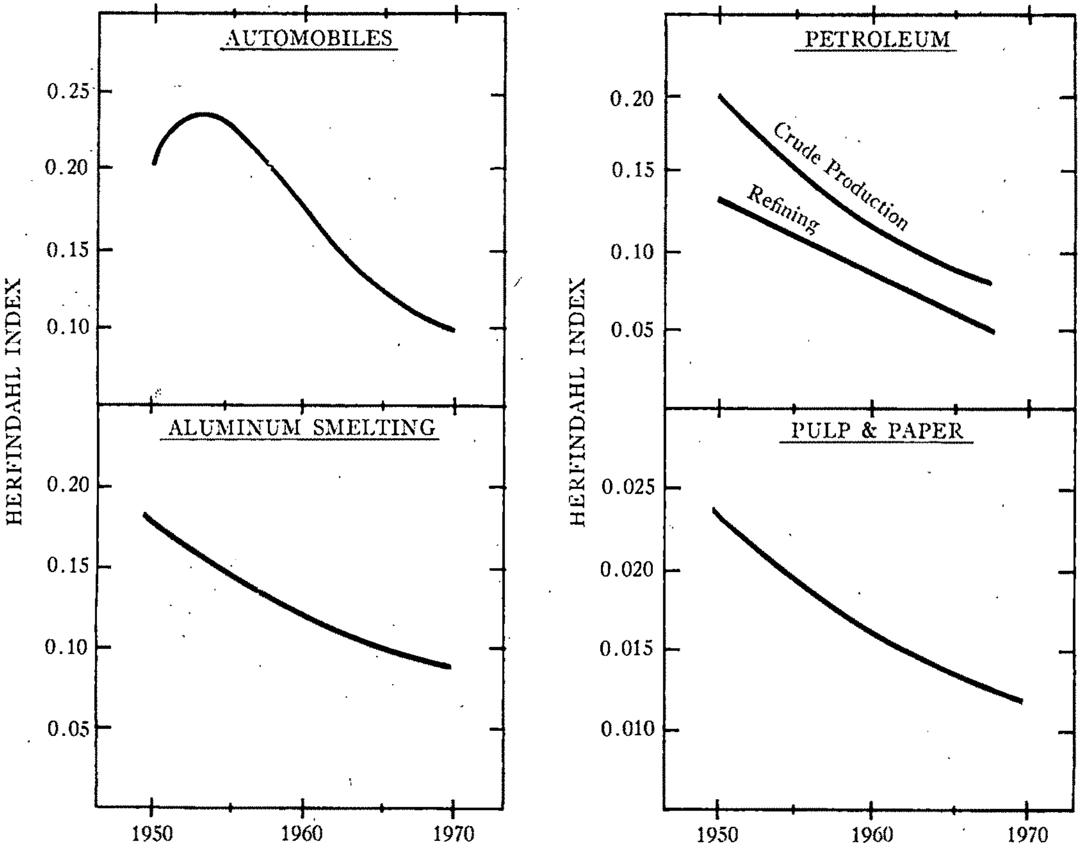


FIGURE 1: HERFINDAHL INDEX OF CONCENTRATION OF WORLD PRODUCTION IN FOUR MATURE OLIGOPOLIES, 1950–1970*

Source: Harvard Multinational Enterprise Study.

* Production of communist countries and other countries with state monopolies is omitted.

patents, and magnified by trademarks and advertising.

Nonetheless, the interpenetration of national markets, already alluded to in the case of the mature oligopolies, has also occurred in the innovation-based industries. Before World War II, when most chemicals could not yet be regarded as standardized or the industry itself as mature, the leaders could divide world markets by mutual grants of exclusive licenses. For example, Imperial Chemical Industries, Ltd. (*ICI*) turned over its *U.S.* patents in specified fields to du Pont, and du Pont, its patents in most of the British Commonwealth to *ICI*, so that in any national market one of the enterprises controlled all of the legal monopolies that had been issued by the nation to either of them.

Today that pattern has practically disappeared. The big multinational enterprises as a rule maintain control over the patents in the areas in which their subsidiaries operate. Electronic firms like Sony, General Electric, and Philips accordingly are more likely to confront each other in many markets, as are chemical firms like *ICI* and du Pont.

The growth of multinational enterprises in the innovation-based industries suggests a number of corollary points. First, multinational enterprises which develop a patentable invention are likely to obtain patents in more national jurisdictions than national enterprises might do (Leonard Robbins). Second, no single national patent is likely to be quite as important to a multinational enterprise as to a national enterprise. Finally, since multinational enterprises appear prone to penetrate one another's main markets more than national enterprises, the degree of differentiation between different patent-protected products will presumably have been reduced, thereby reducing the monopoly value of the patent.

It would be a mistake to conclude, however, that the problems of competition in the innovation-based industries have declined in all respects. Ordinary run-of-the-mill industrial innovation is a relatively low-level, trial-and-error, cut-and-paste activity that many firms in industrial countries are capable of undertaking (Scherer, Edwin S. Mansfield et al.). Some innovation-based projects, however, are so large, prolonged, or complex as to exceed the capabilities of individual firms: an extraterrestrial launcher, a controlled fusion process, a supersonic transport. Projects of this magnitude are not very common, but they are exceedingly important. In these cases, the multinational enterprise which receives governmental assistance has a very decided advantage, not only in the home market but also abroad, an advantage independent of that bestowed by a patent system (Keith Pavitt and *OECD* 1971).

V. Issues of Policy

Geographical divisions of international markets among national oligopolists have become more difficult to sustain. The problems of international competitiveness that remain, however, seem difficult enough. Global oligopolies exist and more may be in prospect. They operate in ways that place both fact-finding processes and remedial measures beyond the effective reach of many individual countries. As a result, individual countries are likely to avoid any action out of a sense of impotence or a desire not to provoke other countries. This suggests the need to take up once again where Chapter V of the Charter for an International Trade Organization (*ITO*) left off a quarter century ago—to inquire if the world is ready for an international fact-finding, adjudicating, and prescribing institution.

Many things have changed since the project of an *ITO* Charter was abandoned in 1950. On the one hand, the ability of the United States to lead any new initiatives in this field has been considerably reduced. On the other hand, various countries on their own initiative have instituted and enforced policies to restrain monopolistic power. Programs of various sorts are to be found notably in the United Kingdom, Germany, the European Economic Community (D. M. Jacobs, W. Schlieder), and even Japan (Masao Baba, Mitsuo Matsushita).

There are some essential differences between *U.S.* programs and foreign programs, however. The tolerance of other governments for restrictive practices is generally higher. And the concept of a per se violation—an act that is *ipso facto* illegal without regard to its economic consequence—is almost uniquely a *U.S.* idea. But these differences do not preclude an international approach that would be more effective than the unilateral application of national laws (Sigmund Timberg). An already existing web of bilateral agreements on the subject between the European Economic Community and various countries in Europe and a number of similar agreements by the United States with various countries offer some of the principles of an international approach.

The solution to the international problems of competition and monopoly, however, probably will require more than a mechanism for more effective fact finding and adjudication in individual cases. Consider, for instance, the international problems associated with patents. The patent represents an implicit bargain between the sovereign state and the inventor who receives a temporary monopoly in return for disclosing the product of his genius. The trend of international policy so far has been to make it easier for the inventor to stretch his monopoly over large areas: this is the general direction of the European patent project, for instance. I have already suggested that multinational enterprises paradoxically may have reduced the value of the patent, both as a stimulus to invention and as an impediment to competition, by the assiduous exercise of their patent rights. This generalization, however, is more appropriate to some products and some markets than others. In many developing countries, for instance, the size of the market and the restrictions on access to the market may have the effect of vesting the patent with real anticompetitive power.

Individual states are not, of course, obliged to join the international trend to make patent grants easier for the inventor. They have it in their power to limit the scope of the patent grant in order to fit the special needs of their own societies. The case for such limitation appears fairly strong in some developing countries. As the developing countries experiment with the modification of patent grants which they feel appropriate to their needs, some forebearance and understanding on the part of other nations may be needed.

There remains the grave problem, half apparent, half latent, of the products that derive their strength in internal markets from governmental support, a situation commonly found in the aerospace, nuclear, and advanced electronic industries of various nations. The issue is already difficult enough among the advanced countries of capitalist persuasion. It will grow even more serious if the Soviets or the Eastern Europeans join the competitive race. Though easy answers are not immediately apparent, the odds are fairly high that we shall soon be obliged to develop an adequate response.

REFERENCES

M. Baba, "Foreign Affiliated Corporations and Concentration in Japanese Manufacturing Industry," Tokyo Conference on International Economy and Competition Policy, 1973, mimeo.

J. H. Dunning, *The Multinational Enterprise,* New York 1971.

C. D. Edwards, "The Impact of International Cartels on International Trade," *Die Konzentration in der Wirtschaft,* Berlin 1971, 137–157.

E. Hexner, *International Cartels,* Chapel Hill 1946

D. M. Jacobs, "Some Developments and Differences in the Operation of Competition Laws," Tokyo Conf. on Intl. Econ. and Compet. Pol., 1973, mimeo.

F. T. Knickerbocker, *Oligopolistic Reaction and Multinational Enterprise,* Boston 1972.

E. Mansfield et al., *Research and Innovation in the Modern Corporation,* New York 1971.

M. Matsushita, "The Japanese Antimonopoly Act and International Transactions," Tokyo Conf. on Intl. Econ. and Compet. Pol., 1973, mimeo.

D. L. McQueen, "Learning, the Multinational Corporation, and the Further Development of Developed Economies," Tokyo Conf. on Intl. Econ. and Compet. Pol., 1973, mimeo.

K. Pavitt, "Technology in Europe's Future," *Research Policy I,* Brighton, England 1972.

L. Robbins, "Patents and U.S. Foreign Policy," *IDEA,* Winter 1967–1968, *2,* 521–551.

F. M. Scherer, *Industrial Market Structure and Economic Performance,* Chicago 1970.

W. Schlieder, "Competition Policy in the

European Community," Tokyo Conf. on Intl. Econ. and Compet. Pol., 1973, mimeo.

S. Timberg, "An International Antitrust Convention: A Proposal to Harmonize Conflicting National Policies Towards the Multinational Corporation," Tokyo Conf. on Intl. Econ. and Compet. Pol., 1973, mimeo.

Y. Tsurumi, "Japanese Multinational Firms," *J. of World Trade Law,* Jan.–Feb. 1973, *7,* 74–90.

J. W. Vaupel and J. P. Curhan, *The Making of Multinational Enterprise,* Boston 1969.

——— and ———, *The World's Multinational Enterprises,* Boston 1973.

R. Vernon, "Organization as a Scale Factor in the Growth of Firms," in J. W. Markham and G. F. Papanek, eds., *Industrial Organization and Economic Development,* Boston 1970.

———, *Sovereignty at Bay,* New York 1971.

European Economic Community, *Second Report on Competition Policy,* Brussels 1973.

Organization for Economic Cooperation and Development, *Conditions for Success in Technological Innovation,* Paris 1971.

———, *Policy Perspectives for International Trade and Economic Relations,* Paris 1972.

U.S. Dept. of Commerce, *U.S. Investments in Foreign Countries,* Washington 1960.

———, *U.S. Direct Investment Abroad, 1966,* Part II, Group 2, mimeo.

———, *Survey of Current Business,* Washington, Nov. 1972.

Problems and Policies Affecting Labor's Interests

By GEORGE H. HILDEBRAND*

It is tempting to draw a parallel between the spread of American industry south and west after, say, 1830 and the spread of multinational enterprises throughout the world, particularly in the past quarter-century. One calls to mind immediately visions of runaway shops, legions of low-paid new nonunion workers, old high-cost plants in declining regions, and the difficulties of extending collective bargaining and common labor standards to remote groups and places. And to be sure, there are similarities, but also some sharp differences.

Perhaps the best way to begin is to set up a conceptual framework with which to examine these questions. First, we should distinguish between the interests of labor organizations in negotiating and enforcing wages and standards of employment and the interests of individual workers, for some union members have interests that do not always parallel those of their organizations and other workers do not belong to unions at all.

Second, we have to consider these interests in all of their geographical complexity: workers and their organizations in a source country for direct foreign investment capital, such as the United States; in the host countries to which the capital is applied; and third, in countries whose workers and their organizations have economic interests that can be either competitive with or complementary to those in the host countries. Here I find it useful to distinguish four admittedly rough categories: the United States itself, the industrial economies of Western Europe and Japan, the developing industrial economies of Asia—Taiwan, Singapore, Hong Kong, and South Korea—and a grab bag of primarily natural-resource-oriented countries, such as Saudi Arabia, Peru, Chile, and Australia. The only brief for this classification is that it helps one perceive the limitations of the runaway-shop model as a basis for looking at direct investment overseas and realize the obstacles to effective extension of union power beyond the seas.

Third, it is safe to assume that capitalistic firms expand abroad to maximize profits. This motive finds concrete expression in different ways. One is the perennial search for cheaper raw materials, for example, richer deposits of oil and copper. Closely related is the desire to take advantage of lower costs of production and distribution, as it were by substituting foreign investment for export sales of manufactured goods. In some cases this phenomenon may involve Raymond Vernon's cycle of overseas expansion, where there exists a joint motive to cut costs and enlarge total sales to overcome saturation problems at home. However, in other situations the purpose may be oriented mainly to cost-savings, for example, low-cost assembly of television components in Taiwan for reexport to the United States. Then again the motive may be to penetrate a foreign market that is closed to imports by a variety of possible barriers—a demand-oriented expansion that may also involve economies of scale as well (export of components, spreading costs of central management, lower prices for inputs purchased in larger volume).

---

* Maxwell M. Upson Professor of Economics and Industrial Relations, Cornell University.

Finally, the objective may be to secure the profit-yielding advantages of vertical or horizontal integration, for example in the vertical case to get favorable cost differentials for raw materials or to obtain an assured supply or, on occasion, even to shut off competitors' access (borates in 1899, bauxite before 1942). In the horizontal case the goal may be to make possible a strategy of shifting production from struck plants (probably overrated) or to get rid of a troublesome competitor or even to reduce the vigor of import competition back home. But horizontal integration may also involve none of these things and, instead, simply may emerge from expansion abroad to improve the rate of return through larger sales and lower costs.

Let me consider first the impacts of direct foreign investment (*DFI*) upon the interests of workers as individuals. Wherever the investment goes, of course, the overall effect is to increase the stock of capital relative to the labor force, increasing employment and, depending upon elasticity of supply of industrial labor, real wages. In relatively undeveloped raw materials countries, the accompanying infusion of advanced technology and managerial skills can make a strong and continuing contribution to the training and upgrading of the local labor force with beneficial spin-off effects for the whole economy. In time, serious political issues may emerge from this invasion by foreign capital and its representatives, whether it be from capitalist or communist countries, and there exist cases of both types. However, I am limiting myself here to purely economic effects.

Of somewhat greater interest, I think, is the effect of the process upon individual workers in the source country, in this instance the United States. The freedom both of capital and of commodities to move internationally operates continuously to shift the participating economies toward specialization according to comparative advantage. Depending upon the case, the home effects may be favorable or unfavorable, complementary or competitive. Funds sent out to develop an oil field in Indonesia will aid the employment and real wages of refining and distribution employees back home, although they may be adverse to the interests of workers in high-cost domestic fields. *U.S.* investments in radio, computer, and television plants in the Asian group or in the Mexican border zone will at least slow the rate of expansion of employment in the collateral domestic branches of these industries and may well reduce it absolutely. However, the lower production costs may also have favorable effects in the other domestic branches of the same industries. The migration of New England textiles southward was a similar process.

Before leaving this topic, let me record my belief that progressive overvaluation of the dollar during the two decades prior to the adoption of floating exchange rates in 1973 has had a great deal to do with both the volume and the nature of *DFI* from the United States. As with the United Kingdom in the 1920's, the effect was to impose a cost-scissors upon domestic industries and their workers, discouraging exports and attracting imports. At the same time, fixed dollar parities were a standing invitation to send capital abroad in place of goods and services. As the whole process went on and its consequences worsened, we began to hear that the United States would eventually be driven out of hard goods production and would become an exporter primarily of computers and managerial talent.

I think we are now able to see matters in better perspective. The evolution of our export and import patterns in these years reflected dollar overvaluation as well as longer run factors of comparative advantage. The cheap labor and runaway shop arguments will continue to be with us, but the forces provoking them should be less severe.,

Let me turn now to the impacts of *DFI* upon unions, using a three-way geographic classification: host country, source country, and "third party" countries. As a general rule we may follow John R. Commons and say that the ability of a trade union to extract gains by bargaining with employers depends ultimately upon the completeness of its control over the relevant product market. "Taking wages out of competition" means eliminating competing uncontrolled sources of product supply, so that differences in wages and benefits do not yield differences in unit costs and that strike strategy may be fully effective.

If the host country for *DFI* is, for example, an industrial country in Western Europe, then the infusion of private capital is likely to pose different problems for a trade union from those that would emerge, say, in Zambia or Saudi Arabia or within a country of the Asian group. In the European case, if the capital is used to acquire an existing firm, the immediate problems are to control the labor policy of the new owners and very possibly to cope with a multiplant technique for combatting strikes. If, instead, the investment is directed to a new plant, the basic problems are likely to be the same, plus the added one of organizing the new work force.

Continuing with the United Kingdom-West European setting, *DFI* in either form in any one of these countries is likely to make the problem of product substitution far more acute all around. Within the host country, the unions will be compelled to look to their fellow organizations for international cooperation against competing employers in the industry. Likewise, these organizations will be furnished with the same incentive. Here there would seem to be three main choices for policy. One is to try to extend the reach of organization, to form genuine international unions to achieve a common policy as regards negotiations, strikes, settlements, and wage and employment standards. Short of this difficult goal, the affected organizations might attempt to form bargaining and strike coalitions, as it were to achieve horizontal integration of the relevant labor markets for both offensive and defensive purposes. Third, they might seek the establishment of "fair" international labor standards for wages, hours, and working conditions, through bodies such as the International Labour Organisation. With this last approach, the problem of what Arnold Weber calls the "mediation of interests" of diverse worker-groups becomes even more formidable.

Consider, next, the case where *DFI* is directed to light manufacturing and assembly in the so-called Asian group of developing countries. Here technology, skill requirements, and predominance of final sales in the United States all enforce a certain commonality and substitutability as among plants within the group. In consequence there is some similarity with the West European case. However, there are differences: unionism is much more unevenly distributed. Where found, the organizations are local and tend to be weak. Distances among competing sources are great, there are ethnic and linguistic barriers, and fraternal ties are usually not close. Taken together, these factors do not make for near-term cohesive action through internationalization of unions, international coalitions, or pursuit of common labor standards. Thus we can conclude tentatively that, whereas in the United Kingdom-Western European region the forces making for cohesive labor policies are strong and growing, dispersion and fragmentation now characterize the Asian group.

Last among the receiving countries for *DFI* are those in the natural-resource-oriented group. These nations are scattered far and wide over the globe. Formally, of course, the country suppliers of petroleum, or copper, or iron ore, for example, are interdependent, but unification of policies on an international scale, where it has occurred at all, has been the consequence of cooperative action by states, not of trade union bodies. On the one side, physical distances are great while competitive disturbances from opening up a new deposit are small. On the other, where independent unionism is found at all, it has neither the power nor the incentive to strive for international lines of action. Everything favors an inward-looking concern for organizing and policing the foreign-owned extractive company, and this is where the practical pay-offs are to be had. Put a little differently, the problem for organized labor typically is to capture and exploit the local resource base, not to reduce the possibilities for substitution of foreign for domestic labor as among far distant locations.

This brings us to the impacts of *DFI* for labor unions in a capital-source country such as the United States. With unimportant exceptions, the outflow of capital to develop new sources of raw materials has not constituted a problem for union policy. Rather, the problem—viewed here as a composite of wages, jobs, and bargaining effectiveness—has asserted itself in two main ways: the loss of export markets, when investment abroad has been substituted, and the substitution of imports for domestic production. From the union standpoint, loss of export sales hurts because of the old principle that capital is more mobile than labor. The same handicap has asserted itself probably even more strongly through import substitution. Impacts here have been felt in apparel, specialty fabrics, footwear, automobiles, fabricated steel, calculators, typewriters, radios, and television sets. To be sure, the effects have been diffuse. In some cases the displacement of workers in particular places has been concealed by overall growth of the industry. In others, the problem appears as slackened growth rather than absolute decline. But for a labor union, significant specific job losses, together with a failure of membership to grow, are serious weakening agents as regards organizational effectiveness.

Interestingly enough, the principal source of these problems seems to be the Asian group and the "twin plants" problem along the Mexican border. I have neither time nor opportunity to examine the statistical evidence, hence must confine myself to a few observations.

One of them is that the products involved require large amounts of unskilled and semiskilled labor. This means a high elasticity of substitution as between the domestic and foreign labor groups involved—always a difficult problem for a labor union. Another consideration is that the competing foreign sources are very unevenly unionized—a version of the open-shop problem. Still another is that all of these countries tend to fit W. Arthur Lewis' model—they have excess supplies of labor, and these exert significant downward pressure against wage increases. This long-distance factor cripples the bargaining effectiveness of domestic unions in the same field.

And finally, while it is cold comfort at best, I suspect that the chief explanation for this phenomenon of sharply increased product-market competition from the Asian group has been simultaneous overvaluation of the dollar and deliberate undervaluation of the yen. Overvaluation limited exportation while promoting a soaring expansion of imports. So far as the United States is concerned, undervaluation made Japan a particularly formidable competitor. How, then, was a domestic producer of, say, television sets to endure against the mortal thrust of Japanese competition? So long as the dollar was firmly pegged and official policy favored liberalized trade, the only way to survive was to have light fabrication and assembly performed in countries that could put the Japanese producers in the same squeeze.

In my view, these are the main situations that have made *DFI* an adverse force for the American unions, a force deriving more from Asian than from European sources.

Obviously one kind of response would be the internationalization of unionism itself—directly or through coalitions. Alternatively, unions could seek "fair" international labor standards. But, with the notable exception of the United Auto Workers and the Japanese auto unions, expansion outward has not been tried. If anything, the American labor movement in recent years has been reducing its ties to its foreign counterparts. As for international economic relations, the line taken has been to support the Burke-Hartke trade bill in all of its intricate protectionist apparatus—import quotas, restrictions upon *DFI*, and restraints upon the diffusion of American technology abroad.

The explanation seems to be that the American unions, like their sister organizations abroad, orient their policies to the interests of their incumbent employed members. Until the last two decades, foreign trade and international movements of private capital offered no significant threat to the position of incumbent jobholders in the organized sections of American industry. With the progressive overvaluation of the dollar and the enormous outflow of private capital, now exceeding $90 billion in cumulative total, certain *U.S.* unions began to feel the pinch—in slackened growth of their industries, in specific job losses, and in diminished bargaining effectiveness (higher elasticity of derived demand for production labor).

As these pressures mounted in the 1960's, they began to impose some hard policy choices upon the American labor movement. Extension of unionism beyond Canadian affiliates to places such as Taiwan or Hong Kong in the Orient, or to Italy and West Germany in Europe, was a manifest impracticality: the organizational requirements were too difficult while the necessary incentive on the side of the prospective foreign members simply was not there. As a second-best alternative, international coalitions offered the American unions even less advantages while the organizational problems were equally difficult. Why, for example, should unions in Hong Kong seek wage parity with their American brethren just to relieve the latter from the competitive rigors of what they would term runaway shops based upon cheap foreign labor? Workers are just not that naive about the negativeness of demand elasticities for their services.

The solution, then, was to turn to measures such as Burke-Hartke because the goal of making secure the interests of incumbent American members seemed easier to obtain given the political power requirements involved.

There was still another alternative, but it was never tried: to press for an end to the Bretton Woods system of fixed parities so that the dollar could attain a competitive position in the world economy. Here was the means to relieve the pressure of imports and to restore the strength of our export industries at one stroke. With the progressive abandonment of administered prices for the dollar in foreign exchange over the last two years, the American unions have gained more for their members than internationalization or Burke-

Hartke could ever be expected to provide them. Perception of this economic truth may well explain the almost total silence now enveloping both of these proposals in the United States.

What, then, can be said regarding the overall significance of private international investment from the standpoint of labor as a factor of production?

First, *DFI* is a vehicle for the diffusion of capital, technology, managerial skills, and occupational training over those regions of the world that are hospitable to the process. As such, what is involved is an increasing international division of labor. In essence the process is no different from westward expansion within the United States itself, except that, at least in its earlier period, that expansion did not entail branch-plant enterprises to the extent that it later did and that development of multinational corporations does today. But this is a question of organizational forms: probably the real significance of the multinational device in this respect is that it simplifies and accelerates the transmission process.

From the standpoint purely of economic welfare, that portion of the world's consumers that falls within the ambit of the process is clearly benefited by its extension. As these consumers are preponderantly workers, the latter are beneficiaries of the increased efficiency in resource use that *DFI* makes possible.

Second, the consequential effect is a widening of the product market and of the number of competing sources of supply effectively within it, although the multinational device works to hold down the number of independent suppliers at the same time that the market itself is widening and expanding.

From the labor standpoint, there are those who gain and those who lose. All workers gain as consumers, most of them as producers, and some lose through the displacement effects of broader competition. Because the unions speak mainly for incumbents on the job and not for consumers or for newly employed workers in lands often far away, the dominant theme in their reaction is hostility to both *DFI* and liberalized international trade. Yet there is reason still to accept what Alfred Marshall taught us more than half a century ago: that on the whole the relationship of capital to labor is far more complementary than it is competitive. Increased employment and higher real wages are the main consequences to be looked for. To search, therefore, for ways to obstruct the export of capital is to look for inefficient and inequitable methods to redistribute income, paradoxical as this may seem.

Looking far ahead, however, the continued spread of the multinationals does presage a shift in the power balance as between the free world's employers and its labor organizations, an emerging gap that the sovereign state, acting alone, seems unable to fill. Indeed, the situation in rough ways is a replica of the United States in the 1920's. The business enterprise has proved fully capable of extending its organizational reach literally all over the world. The world's labor movements have not demonstrated anywhere near as much flexibility and adaptability. As *DFI* proceeds, this gap in the power balance will increase. It is difficult to believe that it will proceed unchecked and unchallenged from the labor side over the decades to come.

TRENDS, PROSPECTS AND POLICIES FOR EAST-WEST TRADE

Prospects and Policy for U.S.-Soviet Trade

By GREGORY GROSSMAN*

The word "prospects" in the title leads me to interpret my assignment as a forward-looking one, a bit of economic futurology. Unfortunately, the economist's tools are of very limited use in this case, especially at this moment in history.

Take the mere question of the level of trade. Simple mechanical extrapolation will certainly not do. Consider that during 1965–1971, *U.S.* imports from the USSR averaged only $53 million per year, with a mild uptrend over the seven years. In 1972 they went up to $95 million, and in the first ten months of 1973 they doubled to an annual rate of $190 million. Even so, they are tiny in relation to the total trade of either country, and therefore intuitively contain a considerable potential for growth. *U.S.* exports rose from $45 million in 1965 to $119 million in 1970 and to $161 million in 1971, then jumped to $546 million in 1972 and to an annual rate of over $1.26 billion in January–October 1973. The 1972 and 1973 figures comprise mostly the extraordinary—but not entirely unrepeatable—grain shipments. It is obvious that the "normal" future levels of *U.S.*-Soviet trade cannot yet be discerned from the recent data.

This much is reflected in the extremely wide range of estimates prepared by competent observers. Since *U.S.* imports from the Soviet Union are likely to remain far smaller than exports (certainly through the 1970's, before large amounts of gas and/or oil from the much publicized joint projects can start arriving on these shores), a major determinant of the level of *U.S.* exports to the USSR for the remainder of the 1970's is the availability of credit, together with the commercial and political feasibility of the giant projects. As the House has just adopted the Vanik Amendment, which would severely restrict credit to the USSR, no apology is necessary for the opaqueness of the crystal ball at this moment.

What economic principles—the political ones will be mentioned later—shape the volume, composition, and direction of Soviet trade with the West, broadly speaking? We will be brief (for extensive treatment see such classic sources as F. L. Pryor, P. J. D. Wiles and Alan Brown). First, the difference in economic systems is usually mentioned as a *negative* factor, one that tends to reduce the volume of East-West trade, and it doubtless has done so. We may cite such things as the mutual suspicion between capitalists and communists, and, on a more operational level, the cumbersomeness of the state trading organizations of the Eastern countries and the lack of direct contact between producers (users) in the East and users (producers) in the West. Not to be overlooked is the constriction of the volume of goods available for export from eastern countries caused by the heavy pressure of domestic requirements and the low marketability of Eastern products in the West for reasons of quality and the like—both of course heavily influenced by systemic factors. And yet, systemic differences can also augment East-West trade. Surely, the large importation of technology by the East from the West at this time, occasioned by the so-called technology gap, should also be in some degree credited to systemic contrasts in regard to the efficacy of research and development and innovation.

---

* University of California, Berkeley.

Systemic differences between East and West may also have a considerable trade-generating effect insofar as they result in disparities in relative prices in the East and West, either through their effect on real variables (e.g., supply of specific factors) or through peculiarities in accounting, calculation, and pricing methods. This brings us to the matter of comparative advantage and its role as a determinant of *U.S.*-Soviet trade.

Perhaps we should rather speak of "pecuniary advantage," given that in both countries actual prices deviate widely from ideal prices of economic theory, whose values we do not know. To the extent that trade between the two countries will be determined by the prospect of pecuniary gain from individual transactions, it will be guided by the actual prices confronting *U.S.* businessmen and Soviet traders. Surely, the trading decisions on the *U.S.* side will continue to be made in the foreseeable future by private businesses and will therefore be price-guided, though probably not without considerable intervention on the part of the *U.S.* government (export controls, subsidies, price support, and the like). We need only recall the role of *U.S.* government subsidy and credit in the billion-dollar-plus sale of grain in 1972. But it takes two to transact; hence the volume and composition of *U.S.*-Soviet trade is equally in the hands of Soviet authorities and traders. To what extent are they likely to be price-guided?

In one important respect we can presume them to be generally price-guided: in choosing American sellers for their imports and buyers for their exportables in preference to sellers and buyers in other noncommunist countries. (Not always though—political reasons may play a perceptible role, both positive and negative, in choosing between American and other transactors, or in deciding to import or export at all.) Moreover, the Soviet Union is now the only East European Comecon country to maintain an almost complete separation between internal wholesale and external prices of importables and exportables (see N. Mitrofanova), which will distort internal decisions relating to both kinds of goods. Beyond this, alas, we simply do not know enough of the process of decision making regarding the volume, composition, and direction of Soviet foreign trade to be able to predict the extent to which such decisions have been or will be price-guided. Surely, autarkic aims must still weigh quite heavily; how otherwise explain the continuing strenuous Soviet efforts toward self-sufficiency in animal proteins and corresponding inputs (feed grain) in the face of rapidly increasing domestic demand and (presumably) high and steeply rising incremental costs? To complicate matters, autarky in the long run may mean more imports of technology and equipment in the nearer term—"imports to end imports"—and of course correspondingly larger exports to pay for them. Moreover, there is a still longer run during which domestic priorities and world technology change sufficiently to negate most of the importance of the autarky that might have already been achieved by dint of great efforts, requiring a new round of strenuous efforts to attain the pot of self-sufficiency at the end of the rainbow. Soviet history provides many examples.

One cannot excape the impression that much of Soviet foreign trade with the West —insofar as it is not primarily motivated by extraeconomic considerations—is to be explained as much by what we may call "gap filling" as by the Soviet quest for pecuniary advantage. By the former, somewhat inelegant phrase we simply mean the attempt to overcome shortages in supplies by importation from abroad and to pay for the augmented imports by exporting such relative surpluses as may be on hand—these decisions being taken primarily in physical terms, i.e., with reference to the standard "material balances." The gap may be seen to be of short or long duration. The need to import is the active cause of trade; exports are the passive, if indispensable, link. (Cf. Pryor, Ch. 4.) A gap in supplies may be anticipated, i.e., "planned" in Soviet parlance; or it may arise only in the course of plan fulfillment owing to planning errors, setbacks of all kinds in the execution of plans, natural causes, and so forth. It can also arise as a result of significant revision of plans and priorities, e.g., the stepped-up importation of consumer goods immediately after Stalin's death for internal political reasons, or the sharp increase in the purchase of western equipment for the chemical industries following Khrushchev's decision in 1958 to "chemicalize" the Soviet economy. In the latter case, as G. Fink shows (p. 108), the average annual value of imports of chemical equipment from all sources jumped tenfold between 1953–57 and 1960–69, and the share of developed Western countries in such imports from almost 0 to 60 percent (the remainder coming from Comecon countries).

If much of Soviet trade with the West is indeed gap filling, then it plays the crucial role of imparting short-run flexibility to the otherwise very rigid Soviet economy, as well as providing greater maneuverability in general over the long run. (Trade with the Comecon offers less flexibility, of course.) Imports connected with short-term gap filling would consist primarily of raw and intermediate producer goods, foodstuffs and other essential consumer goods, incentive goods for sale to consumers, and certain services, such as maritime transport. Long-term gap filling would consist chiefly of capital goods, other investment goods, and technology. The same argument points to the importance of credit: to enable gap-filling imports to take place in the short run before resources can be reallocated to yield the counterpart in exports and to benefit from external financing for long-term projects. The latter has been lately ingeniously cast into the form of "product payback" (or "barter over time") deals, whereby the Western-financed projects are made self-liquidating (bypassing the difficult problem of transfer for debt service) by having the lender take his repayment in the form of the eventual products of the given projects. Incidentally, such deals do point to the exploitation of comparative advantage by the Soviets in that the deal essentially rests on the relative abundance and cheapness of certain natural resources and skilled labor in the USSR. Furthermore, the same projects create additional and economical export capacity to finance gap-filling imports in the future.

At this point it is necessary to look closer at Soviet interest in expanding economic relations with the West in general, and the United States in particular, in the near future, an interest that would appear to be very strong indeed. The economic grounds for this interest are roughly as follows. (1) First is the technology gap, thanks to which relatively small outlay in the West (and in cases of product payback or technology swap, hardly any hard currency cost at all) can often yield large dividends to the Soviet economy—in other words, the stepped-up intake of Western technology will allow the Soviet Union to continue keeping a large part of its inventive and innovative resources in the military sector. (2) The need to turn to more outlying deposits of fuels and other minerals tends to raise the incremental capital-output ratio for at least the near future. (3) The motorization of the country and determined attempts to solve the agricultural problem similarly require large capital outlays in the near term. (4) Not least, the marked slowing down of the growth of the nonagricultural labor force (in the case of industry, from 3.7 percent per year in the sixties to 1.4 percent per year from 1969 to 1973) places additional pressure on both capital formation and technical progress to maintain wonted rates of growth. (Very suggestive are Abram Bergson's illustrative projections.) (5) At the same time, domestic political considerations militate against any significant reduction in the growth of private consumption and social services, as well as (it would seem) any substantial reallocation of resources from defense to capital formation.

A far-reaching, decentralizing economic reform might yield additional resources by giving a one-time boost to static efficiency, but then again it might not and in any case is anathema to the current leadership for power-political reasons. There remains massive assistance from the West in the form of large capital flows, technology transfer, managerial and organizational know-how, and gains from merchandise trade. And potentially the largest assistance can be obtained from the United States, because we have more than anyone else to offer and because more than anyone we stand to take in exchange promises of delivery of that ethereal substance, the *détente*. In a real sense, therefore, our assistance is an alternative to decentralizing reform and to reduction in defense outlay, rather than a stimulus to them.

Indeed, the political aspect cannot be separated from the economic in an assessment of Soviet motives in expanding economic relations with the United States. We have just alluded to the domestic political implications. On the international plane one need only mention the isolation of China, and the (shall we say) bridging over Eastern Europe that bids to reduce both its economic bargaining power and its political importance. But perhaps the greatest political dividend of all expected by the Soviets from a policy of maximal economic involvement with the United States is the creation of powerful vested interests in this country, anxious not only to preserve the lucrative trade but also to safeguard billions of dollars' worth of loans to the USSR, interests whose lobbying could decisively enhance the freedom of maneuver of Soviet foreign policy by restricting that of the American. The syndrome is not novel; it was masterfully analyzed (with regard to the Nazis) over thirty years ago by Albert O. Hirschman, who stressed the "influence effect" of trade.

But what of the *U.S.* interest in greatly enlarged economic relations with the USSR? Here we must distinguish sharply between private gain and national benefit. The private gain for our traders, producers, and financiers is obvious. There should also be some social benefit in the form of usual gains from trade (cf. Franklyn Holzman), despite the relatively modest likely volume. At the same time, in these days of a resuscitated balance of payments, very tight money, overload of our equipment manufacturing industries, and a world crying for food, any very large exports to the USSR, which of necessity would have to be mostly on credit anyway, raise major questions. Nor are the political gains from —so to say—underwriting the Soviet economy for the near future unambiguously positive. The prospect of significant imports of fuels thanks to the development of Soviet resources may be cited in rebuttal. But the fuel would begin to arrive only in the 1980's, at high prices, and after massive commitment of American capital and technology. It may well be more prudent for us to invest comparable resources into the development of domestic self-sufficiency in energy, so as to forestall both vulnerability in energy supplies and possible leverage against us by a powerful long-term debtor.

The chief interest to the United States in greatly enlarged economic relations with the USSR lies in a different direction, namely, by linking them with the normalization of relations between the two superpowers on various political and cultural planes. Given strong Soviet interest in our goods, technology, and capital, there would seem to be an opportunity to use the peaceful instrument of our material wealth in a positive but measured way, and without major detriment to third parties (such as the People's Republic of China), to achieve a more stable world.

REFERENCES

A. **Bergson**, "Soviet Economic Perspectives: Toward a New Growth Model," *Problems of Communism*, Mar.-Apr. 1973, 1-9.

A. A. **Brown**, "Towards a Theory of Centrally-Planned Foreign Trade," in A. A. Brown and E. Neuberger, eds., *International Trade and Central Planning*, Berkeley 1945.

G. **Fink**, "The Quantitative Dependence of Capital Investment in the USSR on Imports of Machinery and Equipment," Osteuropa-Institut München, *Forschungsbericht 1971*, Munich 1971, 87–114.

A. O. **Hirschman**, *National Power and the Structure of Foreign Trade*, Berkeley 1945.

F. **Holzman**, "East-West Trade and Investment Policy Issues: Past and Future," in U.S. Congress, Joint Econ. Comm., *Soviet Economic Prospects for the Seventies*, Washington 1973, 660–69.

N. **Mitrofanova**, *Planovoe khoziaistvo*, 1973, 9, 90–97.

F. L. **Pryor**, *The Communist Foreign Trade System*, Cambridge 1963.

P. J. D. **Wiles**, *Communist International Economics*, New York 1968.

Sino-American Trade Prospects and Policy

By ALEXANDER ECKSTEIN AND BRUCE REYNOLDS*

Between 1950 and mid-1971, the level and even the commodity and geographic composition of China's foreign trade were conditioned by the total embargo on trade and payments between the United States and the People's Republic of China (*PRC*). China's imports were also shaped by more or less stringent controls by all North Atlantic Treaty Organization (*NATO*) countries and Japan. The lifting of the *U.S.* embargo in mid-1971 changed this situation dramatically. In this paper we first analyze some of the factors that shaped China's total trade over the past twenty years and then turn our attention specifically to Sino-American trade in the wake of this dramatic change.

I. Long-term Trends in China's Foreign Trade

China's economic performance in the 1960's stands in marked contrast to that of the 1950's. From the time of economic recovery in 1952 to the peak level of economic activity in 1959, China's exports rose in real terms at an average rate of almost 17 percent a year, while imports rose about 13 percent a year. At the same time, *GNP* grew at an annual rate of about 8 percent. In sharp contrast, when the *GNP* growth rate declined to only about 3 percent a year, foreign trade growth came to a virtually complete halt (see Table 1): exports rose half a percent (in real terms) and imports fell slightly. Thus the swing in export and import growth rates between the 1950's and 1960's was much greater than that in *GNP*.[1]

How can this pattern of trade and growth be explained? We hypothesize that China's economic growth contains a strong "import-led" component, and conversely that growth of China's imports is in part a function of the pace of investment expansion. In the crude model we have in mind, policy measures increase the rate of investment and make the composition of investment goods more technology-intensive. This will then stimulate higher growth in both *GNP* and imports. The stimulus to imports will necessarily exceed that to *GNP* for two reasons. First, investment will grow faster than *GNP*, and imports are functionally related to the level of investment. Second, a shift to technology-intensive investment goods will require that a high proportion of such goods be imported. Furthermore, these imports may raise the productivity of investment, further stimulating *GNP* growth and thus permitting still further rises in investment and imports. This expansionary process eventually encounters domestic savings and balance of payments constraints. As contraction occurs, the growth rate of imports may be expected to fall to that of *GNP*, or below it, for reasons analogous to those adduced above.

This would explain why foreign trade growth was rapid in the 1950's and near zero in the 1960's. For the two decades combined, trade grew at 5 percent a year—somewhat higher for exports and less for imports—that is, perhaps marginally higher than did *GNP*. What does all this suggest for the future course of the Chinese economy and its foreign trade?

On the basis of past indications it would

---

* University of Michigan.

[1] See Eckstein (p. 232, Table 6).

TABLE 1—CHINA'S FOREIGN TRADE, 1952–1972
(In Millions of U.S. Dollars)

	In Current Prices		In Constant 1963 Prices	
Year	Exports	Imports	Exports	Imports
1952	841.2	931.0	764.0	922.0
1953	1,001.3	1,158.2	960.3	1,194.4
1954	1,083.1	1,319.6	951.8	1,377.5
1955	1,298.9	1,380.2	1,224.2	1,427.3
1956	1,564.7	1,540.5	1,493.0	1,534.4
1957	1,563.2	1,488.8	1,481.7	1,427.4
1958	1,861.7	2,004.7	1,861.7	2,004.7
1959	2,159.1	2,119.9	2,239.7	2,145.7
1960	1,950.4	2,049.8	1,908.4	2,091.6
1961	1,544.3	1,513.4	1,556.8	1,545.9
1962	1,557.1	1,207.5	1,620.3	1,238.5
1963	1,611.8	1,230.7	1,611.8	1,230.7
1964	1,811.7	1,489.9	1,742.0	1,453.6
1965	1,984.0	1,941.9	1,952.8	1,879.9
1966	2,199.0	2,070.7	2,143.3	1,948.0
1967	1,962.5	1,995.9	1,946.9	1,882.9
1968	1,916.1	1,800.3	1,893.4	1,716.2
1969	2,004.6	1,837.6	1,896.5	1,698.3
1970	2,011.9	2,189.5	1,830.7	1,849.2
1971	2,314.9	2,253.1	2,091.2	1,837.8
1972	2,885.1	2,769.1	2,404.3	2,066.5

Methods and Sources:

Trade data in current prices were obtained from the following:

For 1952–1962 from Alexander Eckstein, *Communist China's Economic Growth and Foreign Trade*, New York 1966, Tables 4–1 and B–1.

For 1963–1967 from Robert Dernberger, "Prospects for Trade Between China and the United States" in Eckstein, ed., *China Trade Prospects and U.S. Policy*, New York, 1971, Table A.3.

For 1968–1972 compiled by the authors from the official statistics of China's trading partners and from *CIA*, *PRC: International Trade Handbook*, Washington, Oct. 1973.

These trade data were uniformly adjusted so that *PRC* exports are f.o.b. and imports are c.i.f. For Western, Eastern and Southern Europe an adjustment factor of 15 percent was used; for the Western Hemisphere and Sub-Saharan Africa it was 10 percent; for East and South Asia and for Oceania it was 5 percent; for the Soviet Union and Asian communist countries no adjustment was applied on the assumption that most of the trade would move overland with these of China's neighbors.

The trade data in current prices were deflated on the basis of *U.N.* indices of world trading prices with 1963 = 100. *PRC* exports were deflated on the basis of price indices for developing countries by four commodity categories: food, agricultural products other than food, minerals, and manufactured goods. However, for the last category we had to use an "all commodities" index as a proxy since the *U.N.* sources do not provide an *LDC* index for manufactures. For *PRC* imports we used the developed country price indices for these same categories. These price indices were then weighted on the basis of the commodity composition of Chinese exports and imports. The indices were culled from the *U.N. Yearbook of International Trade Statistics*, Table D for various years; commodity composition estimates were compiled by us based on R. Dernberger, "Prospects for Trade Between China and the U.S.," in A. Eckstein, ed., *China Trade Prospects and U.S. Policy*, New York 1971, Tables A–6 and A–7 and *CIA*, *Handbook*, Tables 5 to 9.

be reasonable to assume that, barring internal or external political shocks, the Chinese economy may be expected to grow at a rate of 4 to 5 percent a year. This is suggested not because it happens to be the average for the past two decades, but rather for certain quite clearly discernible structural reasons. It is improbable that the 7 to 8 percent rates of growth achieved in the 1950's can be sustained again over any prolonged period, since they were in part made possible by once-for-all factors operative at that time. More fundamentally, it is most unlikely that Chinese agricultural output will increase faster than 3 percent a year. With about 80 percent of the labor force tied down in the rural areas and with rates of investment already quite high by most standards, it is most improbable that industrial expansion could be so rapid as to push the annual rate of *GNP* growth above a long-term average of 5 percent a year.[2]

There are strong reasons for supposing that the Chinese economy is once more embarked on the kind of growth pattern described earlier. As the economy transits from an average rate of growth of 2 to 3 percent in the 1960's to a 5 percent rate in the 1970's, the rate of investment may be expected to rise. Therefore, for a few years at least, we may expect a more rapid rise in the investment rate than in *GNP*. Finally, the import component of investment will almost certainly increase in the coming years for a number of reasons.

---

[2] If China were to obtain substantial long-term credits, it could, of course, significantly ease these constraints to growth. There is no question that within the last year or two a series of high-level policy decisions have been made to pursue a more active and open international economic policy. However, this does not yet encompass the acceptance of long-term credit, but does provide for medium-term commercial loans with some forms of government backing.

There are very clear indications that the Chinese are investing heavily in a number of high technology projects. After many years of limited plant expansion, a number of new projects are being initiated in the chemical, petrochemical, and petroleum extracting industries, coupled with some expansion in steel capacity and transport equipment. The need to catch up in plant expansion foregone is reinforced by the marked lag in borrowing advanced technology during the sluggish growth, low import, and autarkic policy years of the 1960's. The clearest evidence that the Chinese policy makers are in fact guided by these considerations is provided by the fact that in 1973 the People's Republic entered into contracts for the delivery of complete plants (turnkey projects) from Japan, Western Europe, and the United States valued at between $1.2 and $1.5 billion.

This suggests that Chinese foreign trade may revert to the pattern of the 1950's when it rose at a 15 percent rate. But since, as noted above, it is most improbable that *GNP* will rise as rapidly in the 1970's as in the 1950's, a 10 percent average annual foreign trade growth rate seems more plausible between 1972 and 1980. This would mean that *total turnover in 1980 would be more than twice as high as in 1972.* Moreover, in contrast to the past pattern, we would expect China's imports to rise somewhat faster than its exports. The Chinese will have at their disposal some medium-term credits which will enable them to maintain a trade deficit. Assuming such credits will grow in volume, China will have less need of a trade surplus than when it was repaying Soviet credits between 1955 and 1965.

Under these assumptions, imports might rise to $7.6 billion by 1980 (in 1973 prices) with exports increasing somewhat more slowly. The plausibility of this projection seems to be buttressed by the preliminary estimates for Chinese exports and imports in 1973. Apparently propelled by the new wave of domestic expansion on the one hand and the *PRC*'s more open international economic orientation on the other, total trade in real terms seems to have risen by about 15 percent in 1973, with a much faster increase in imports than in exports and a small trade deficit resulting therefrom.[3]

II. Sino-American Trade Prospects

According to most recent government estimates, our 1973 exports to China will be around $700 million and our imports $60 million. This indicates a doubling in imports over 1972 and an almost twelvefold increase in exports.[4] As a result, the trade imbalance would rise from a 2:1 ratio to almost 12:1. The most fundamental reason for this imbalance is that the United States has a comparative advantage in the exports of certain kinds of products (farm products, transport equipment, and machinery) all of which loom large in China's import bill. The reverse is much less true. Moreover, some marked institutional obstacles hamper China's entry in the *U.S.* market, such as lack of most-favored nation status (*MFN*), the frozen assets problem, and the time required for Chinese exporters to adapt to American marketing techniques. There are no corresponding obstacles now on the export side.

---

[3] According to as yet unpublished *U.S.* government estimates of early December 1973, total *PRC* exports were projected as $3.7 billion and imports as $3.8 billion in current prices. At the time that this paper was written, world export price indices were available only for the first half of 1973. We made some allowance for further price increases in deriving an estimated 15 percent for the real increase in 1973 trade volume. However, the deflator used may be subject to considerable revision.

[4] These are taken from *Current Scene* (p. 10, Table V). They differ quite significantly from U.S. Central Intelligence Agency (*CIA*) data (p. 10), which indicate *U.S.* exports to China of $79 million instead of $60.2 million. This discrepancy arises from the fact that the U.S. Department of Commerce treats *U.S.* exports to China on an f.o.b. basis. The *CIA* views this as *PRC* imports and therefore adds a c.i.f. allowance to the Commerce Department figures.

A. Exports

This year China will have obtained about 18 percent of its imports from the United States as compared to approximately 25 percent from Japan. It is improbable that over the long run the United States will supply more than 20 percent of *PRC* imports or that Japan and the United States combined will provide more than 50 percent. This means that *U.S.* exports to China would be about $1.5 billion in 1980, about twice the current level. It would imply that total *PRC* imports and purchases from the United States would rise at about the same rate, with the latter moving ahead a little faster than the former. What about the commodity composition of these exports?

It is often assumed that *PRC* imports of *U.S.* agricultural products are purely temporary, a function of a poor 1972 harvest. However, this is most doubtful. All current indications point to at least the same level of farm exports from the United States to China in 1974 as in 1973. But beyond this, China has been a net importer of grain since 1961, even following good harvest years, probably in part to relieve the burden on its domestic distribution and transport system and thus ease the strain of supplying the coastal cities. This policy might be expected to continue through the rest of this decade although the actual quantities demanded may fluctuate in response to domestic harvest conditions. Grain imports from Australia and Canada may be expected to remain reasonably stable, while our larger production makes us better able to meet rapid increases in the Chinese demand for grain.

The long-term outlook for Chinese imports of cotton from the United States is also quite favorable. Cotton imports may be expected to rise since the textile industry is expanding in response to increasing population, a gradually rising standard of living, and growing pressure from planners to expand China's textile exports. Given the limited opportunities for expanding cotton acreage and even with some allowance for rising yields, China will almost certainly have to depend on increasing cotton imports—at least until the end of this decade by which time synthetic fiber production will have increased significantly, based on the new plant capacity on order now in Japan and Western Europe.

It may, therefore, not be unreasonable to expect that over the next five years or so *U.S.* agricultural exports to China will at least stay at current levels while the exports of transport equipment, machinery, and various high-technology products will rise. To derive a somewhat clearer picture of a projected commodity composition for *U.S.* exports to China in 1980, we calculated for the 1959 to 1969 period, at the one-digit Standard-International Trade Classification (*SITC*) level (0–9), the commodity distribution of total *U.S.* exports, *U.S.* exports to Asian less-developed countries (*LDC*'s), total imports of those *LDC*'s, total Chinese imports, Chinese imports from developed Organization for Economic Cooperation and Development (*OECD*) member countries and total exports of those *OECD* countries.

TABLE 2—COMMODITY COMPOSITION OF U.S. EXPORTS TO AND IMPORTS FROM CHINA IN 1973 AND PROJECTED FOR 1980 IN PERCENT

Commodity Category		Exports		Imports	
		1973[a]	1980	1973[a]	1980
Food and Live Animals	(0)	62.2	40	9.3	15
Beverages and Tobacco	(1)	0.2	—	0.6	—
Crude Materials, Inedible, Except Fuels	(2)	25.5	10–15	25.1	26
Mineral Fuels, Lubricants, and Related Materials	(3)	0.0	—	0.6	—
Animal and Vegetable Oils and Fats	(4)	4.1	—	1.3	3
Chemicals	(5)	0.6	5–10	12.5	—
Manufactured Goods, Classified Chiefly by Material	(6)	1.1	5–10	30.1	31
Machinery and Transport Equipment	(7)	6.0	30	0.6	—
Miscellaneous Manufactured Articles	(8)	0.1	5	18.1	15
Commodities Not Classified According to Kind	(9)	0.0		1.3	10[b]
Total (in millions of 1973 dollars)		482.6	1,500	50.1	325–390

[a] Actual, Jan.–Oct.

[b] This really encompasses "all other," i.e., elements of all omitted categories.

Sources: 1973 data, U.S. Dept. of Commerce.

1980 projections based on data in *OECD*, *Trade Statistics*, Series C, *I* and *II*, various years; *U.N. Yearbook of International Trade Statistics*, New York, various years; *U.N. Monthly Bulletin of Statistics*, Sept. 1973; U.S. Dept. of Commerce, *Survey of Current Business*, Aug. 1973. For the method used see text.

We used these different series to project shares that seemed to us reasonable based on the past behavior of *U.S.* exports to East Asia and India and the imports of the *PRC* and of these countries. As may be seen from Table 2, we project a decrease in the *relative* importance of foods and a significant rise in machinery and transport equipment between 1973 and 1980—from about 62 to 40 percent in the first case and from 6 to 30 percent in the second. This would represent a 25 percent rise in farm exports and a four-and-one-half-fold increase in the shipment of investment goods.

B. Imports

U.S. exports to China may be considered to have reached a more or less "normal" level in 1973. However, the situation is quite different in respect to *U.S.* imports from China, where there are not only major learning problems but some significant trade barriers as well, such as the frozen assets problem and lack of *MFN* status for China. It is far from clear how important absence of *MFN* in and of itself is as a barrier to imports from China. In a recent study it is estimated that granting of *MFN* would raise our imports from the *PRC* by 16 percent, which applied to the 1973 data would mean an increase from a projected $60 to $70 million.[5] Even if this were a serious underestimate, it would seem that *MFN* could not make a significant dent in the large trade imbalance of 1973. Of course, *MFN* also has great symbolic significance and may be quite important in political and psychological terms, which in turn may have some effect on levels of trade. Even if *MFN* alone is not a major barrier, the other institutional obstacles are real. We proceed on the assumption that by 1980 all these barriers will be removed.

---

[5] See Steven C. Haas.

In projecting Chinese exports to the United States, we started by estimating the future rate of growth of our total imports. The commodity composition of total *U.S.* imports in *SITC* categories 0, 2, 4, 6, and 8 was then projected based on trends for a twenty-year period (1952 to mid–1973). These categories encompassed 90 percent of China's exports. We next assumed that there are no, or only minimal, institutional barriers to Western Europe's imports from China and that in the future any barriers remaining in Sino-American trade will not differ from those of Western Europe in the past. On this basis we hypothesized that a portion of Chinese exports currently flowing to Western Europe would flow to the *U.S.* market, displacing other Asian exports. The resulting pattern of *U.S.* imports was then calculated based on the following four assumptions: (1) that all Asia's share of the *U.S.* import market would not increase; (2) that the same would be true for all Asian exports in Europe; (3) that (a) the proportion of all Chinese exports and (b) the proportion of all other Asian countries' exports sent to Europe and the United States combined would not increase; and (4) that China's share of all Asian exports would be the same in the *U.S.* and European markets.

Projections based on past trends are necessarily hazardous since they cannot take account of changes in domestic economic structures and patterns of trade. Yet the results seem plausible. As shown by Table 2, *U.S.* imports from China are projected at $300 to $400 million (in 1973 prices) for 1980 depending on whether a 5 percent or 8 percent rate of import growth is assumed over the next seven years. It is clear that these projections still leave China's trade with the United States quite unbalanced, with an export-import ratio of about 4:1 or 5:1. However, this would represent a marked improvement over the current situation. This relationship and, indeed, the whole character of Sino-American trade could markedly change if China were to become a large-scale exporter of oil and oil products during this period. The *PRC* is expected to ship one million tons of crude oil to Japan in 1973 and there are discussions under way for the possible delivery of three million tons in 1974. There are also indications from trade sources that the Chinese have started some offshore drilling in shallow waters. Should China's petroleum production continue to grow at recent rates, even allowing for a rise in domestic requirements, the *PRC* by 1980 could become a sizable oil exporter.

REFERENCES

A. Eckstein, "Economic Growth and Change in China: A Twenty-Year Perspective," *The China Quart.*, Apr.–June 1973.

S. C. Haas, *Impact of MFN on U.S. Imports from the PRC*, Office of East-West Trade, U.S. Dept. of State, Washington, Aug. 17, 1973.

Current Scene, "China's Foreign Trade," Hong Kong, Oct. 1973, *11*.

U.S. Central Intelligence Agency, *People's Republic of China: International Trade Handbook*, Washington, Oct. 1973.

Prospects for U.S.-East European Trade

By Alan A. Brown, Paul Marer, and Egon Neuberger*

What are the prospects for United States-East Europe (*EE*) trade, and how might possible policy changes affect the outlook? While we do not attempt to discuss the full complexities of these questions in the limited space available to us, we do touch on a number of relevant considerations in the hope that our discussion will be an appetizer to the main course, to be served by us in a forthcoming study for the Joint Economic Committee. In this paper, after presenting some factual background, we examine problems associated with forecasting *U.S.-EE* trade. While forecasts vary, all indicate that the United States will almost certainly have a large trade surplus with *EE*. This leads us to consider the ways in which *EE* might finance its prospective deficit and to report on a survey of the views of *EE* individuals and institutions on the future of trade with the United States.

I. Factual Background

While the combined populations and *GNPs* of the six members of the Council for Mutual Economic Assistance (*CEMA*)—Bulgaria, Czechoslovakia, East Germany, Hungary, Poland, and Romania—are much smaller than those of the USSR, they conduct much more trade with the West. In 1972 *EE*'s total trade turnover (exports and imports) was about $44 billion and its turnover with Organization for Economic Cooperation and Development (*OECD*) member countries was $11.5 billion vs. $29 billion and $6.4 billion, respectively, for the USSR. Until the large 1972–73 grain deal and the significant rise in *U.S.* machinery sales to the USSR, the *EE* countries combined were also more important trading partners of the United States than was the Soviet Union, as shown in Table 1.

Table 1 indicates (1) a very sharp break in the series on United States-USSR trade in 1972 and a somewhat less sharp but still significant break in the *U.S.-EE* trade series in 1973; (2) the very low levels of *U.S.* trade with both the Soviet Union and *EE*, in relation to total trade of both partners; and (3) the continued *U.S.* trade surplus, which has recently increased dramatically due to the increased availability of credits from the United States. (Of the projected 1973 $1.4 billion *U.S.* trade surplus with the USSR and *EE* combined, roughly $1 billion is being financed by governmental and private credits, as compared to $138 million of credits in 1972, and only $28 million in 1971.)

II. Forecasting U.S. Trade

Problems of Methodology and Measurement. To develop a satisfactory forecast of future trends in *U.S.-EE* trade, one would need a dynamic, general-equilibrium model, including explicit treatment of uncertainty on such issues as future world market conditions and governmental policies in East and West, possible reactions of our Western competitors to any of the hypothetical changes in *U.S.* trading policies, and developments within *CEMA*. Worse still, data on certain key variables, such as invisibles and capital transactions, are not available. Moreover, available data from

---

* University of Windsor, Indiana University, and State University of New York, Stony Brook, respectively. We gratefully acknowledge support by the International Development Research Center of Indiana University. We are associated with the Center as research scholars.

Table 1—Trade of United States with Eastern Europe and USSR

(In Millions of Current Dollars)

	1960–63 Average		1964–67 Average		1968–71 Average		1971		1972		1973 Estimated	
Trade Partner	Exports	Imports	Exports	Imports	Exports	Imports	Exports	Imports	Exports	Imports	Exports	Imports
East Europe	122	59	144	109	189	151	221	167	271	226	600	300
USSR	32	21	73	39	110	60	161	57	547	95	1,300	200

Source: 1960–1970: *OECD, Statistics of Foreign Trade,* Series A, annual issues; 1971–1973: U.S. Dept. of Commerce' Bureau of East-West Trade, "U.S. Trade Status with Socialist Countries," Nov. 30, 1973.

different sources on the same variable—"mirror statistics"—often differ. Large relative discrepancies can be found between various series on *U.S.-EE* trade (see Table 2); that this is not a quirk of *U.S.* statistics is indicated by the data on Soviet and *EE* trade with *OECD* countries, where mirror differences are also very large and not readily accountable for by such factors as transport costs or the method of showing "provenance" of imports and destination of exports. For example, Soviet statistics on the USSR's 1960–71 cumulative trade balance show a Soviet deficit of $915 million, while *OECD* data indicate a surplus of $3.3 billion. These and similar problems are elaborated elsewhere (Marer, Appendices D and E). But even beyond the known data problems, there are also such questions as how much of the recent Soviet grain purchases from the United States found their final destination in *EE.* (We know that some grain was unloaded in Polish ports.) Similarly, some of *EE* purchases from West Europe (*WE*) are made from subsidiaries of multinational corporations owned largely by *U.S.* interests.

Even if all the data were available, it would be difficult to determine, for forecasting purposes, the proper specification of the functional relationship when there

Table 2—Cumulative 1960–1972 Trade Balance Between the United States and East European Countries: U.S. Data Compared with East European Statistics

(In Millions of Current Dollars)

U.S. Trade Partner	Cumulative 1960–1972 Trade Balance: *U.S.* Source	Cumulative 1960–1972 Trade Balance: *EE* Source	Difference as Percent of *U.S.* Exports
Bulgaria	$ 22.6	$−3.5	58
Czechoslovakia	13.4	27.4	5
East Germany	150.0	205.5	24
Hungary	105.6	160.9	35
Poland	95.9	12.7	8
Romania	208.0	196.5	4
East Europe	595.5	599.5	0

Source: OECD, Statistics of Foreign Trade, Series A, annual issues, and official East European publications.

are sharp breaks in the series, if only because the relationship itself may be changing. It is dangerous to extrapolate the earlier relationship, foolhardy to assume a continuation of the most recent one, and worse yet merely to fit a single regression to the whole series and to ignore the discontinuity. The problem is further complicated by the existence of fluctuations in East-West trade. A preliminary econometric study determined that during 1955–69 *OECD* exports to *CEMA* and *OECD* nonagricultural imports from *CEMA* fluctuated more than total *OECD* exports and imports due to both demand and supply conditions and that the fluctuations tended to have some cyclical elements (Jan Stankovsky). The separation of cyclical movements from random disturbances is most difficult, even if the appropriate functional relationship is known so that the series can be purged of trend elements (Neuberger).

Existing Forecasts. Despite the problems bedevilling a would-be forecaster, some intrepid souls on the staff of the U.S. Department of Commerce's Bureau of East-West Trade (Erast Borissoff and Stephen Sind) have attempted to predict *U.S.-EE* and U.S.-USSR trade trends. (See Table 3.) Their methodology follows closely that

TABLE 3—UNITED STATES EXPORTS TO EASTERN EUROPE AND THE SOVIET UNION, ACTUAL FOR 1971–1973 AND PROJECTED FOR 1978

(In Millions of Current Dollars)

	Actual Exports					
Trade Partner	1971	1972	1973[a] (est.)	Estimate	1978 If Trade is "Maintained"[b]	1978 If Trade is "Normalized"[c]
Eastern Europe	221	271	600	High	607	1,550
				Middle	465	1,239
				Low	356	888
USSR[d]	161	547	1,300	High	357	1,051
				Middle	—	—
				Low	232	684

[a] Estimated on the basis of actual trade for January–October 1973.

[b] Assumes that the commodity composition would remain the same as in 1970 and that the *U.S.* percentage share of fifteen industrialized countries' exports to individual *EE* countries remains fixed.

[c] Assumes elimination of discriminatory export and credit controls; that for *SITC* categories 0–4 and 9, exports would be the same as if trade is "maintained"; and that for *SITC* categories 5–8, exports would be determined as outlined below.

[d] Actual includes grain.

Source: Actual: as cited for Table 1; 1978: Borissoff and Sind.

Note: The procedure underlying the forecasts is: (1) to exclude the abnormal 1972–73 grain deal, which explains why the forecasts seem low in comparison to actual 1972–73 *U.S.* exports to the USSR; (2) to present two principal variants: with and without relaxation of discriminatory controls, and, within each variant, high, middle, and low estimates, representing the fitting of different curves to different time periods; and (3) to make the assumption that the "normalized" *U.S.* share in manufactured goods imports by *EE* and the Soviet Union would be the same as the *U.S.* share in imports of these goods by advanced Western countries, thus assuming that the United States will have the same competitive position in *EE* markets as in Western markets. "Normalization" includes abolition of all *U.S.* export controls that are more severe than those of principal competitors and of the limitations on Export-Import Bank credit guarantees.

of Thomas A. Wolf (1972, 1973a).[1] As a rough approximation, *U.S.* exports to *EE* could double in five years from the relatively high 1973 levels, to about $1.2 billion, according to the "normalized," middle-level forecast.

To be comprehensive, these forecasts would have to incorporate additional assumptions, particularly (since the methodology is based on the *U.S.* share of total Western exports to *EE*) about trends in total *EE* imports and in the share of these imports imported from the West. Moreover, as price changes have become more important than they were during the 1960's, attention would have to be paid to differences between the value and the real volume of trade.

In any case, a key factor influencing the level of *U.S.* exports to *EE* is the ability of *EE* to finance the large expected trade deficit resulting from the increases in projected imports from the United States. What options are open to *EE* if it seeks to finance this prospective deficit? Will *EE*'s choice of options—or the likely imbalance itself—influence the feasible level of *U.S.* exports and therefore introduce additional uncertainty into forecasts of them?

III. Options to Finance Imports from the United States

MFN or Not? Granting of most-favored nation status (*MFN*) and removal of some nontariff barriers (*NTBs*) may have substantial impacts on the ability of *EE* to provide the type and quantity of exports salable in the United States. A. Malish, a staff member of the U.S. Tariff Commission, concludes that lack of *MFN* has hurt *EE* much more than the USSR and the industrial countries of *EE* much more than the less-advanced countries. Thus, the dynamic effect of lack of *MFN* is that, as countries industrialize, they suffer more from this type of discrimination. Another study (Wolf 1973b) concluded that the rise in Polish exports was not systematically related to the theoretical gain in price competitiveness caused by reduction of the duties to *MFN* levels in 1960, when Poland became the first *CEMA* country to receive *MFN*.

Western experts tend to doubt that the granting of *MFN* would have a significant impact on the ability of *EE* countries to increase their sales of manufactured goods to the United States. On the other hand, all of our *EE* respondents placed great emphasis on *MFN*, arguing that, in addition to the benefit of lower tariffs on those goods where demand and supply elasticities are high, there would also be very important political and psychological benefits. According to them, the *U.S.* market is very large, sophisticated, and segmented; penetrating it will require a heavy investment in market research, advertising, and other selling costs, as well as the allocation of a large portion of output. The absence of an acceptable political framework, symbolized by lack of *MFN*, reduces greatly the willingness of planners to bear the high costs of entering the *U.S.* market with a wide range of manufactured products. Thus, they stress the importance of the dynamic effects of goodwill created by the granting of *MFN*, which they see as a precondition to changes in existing trade patterns.

Even if *EE* obtained *MFN*, its ability to

---

[1] Wolf analyzed the effect on *U.S.* exports of removal of unilateral export controls and of prohibition on Export-Import Bank (*EXIMBANK*) support and showed that in 1968 the former would have added $342 million to the actual level of $217 million of *U.S.* exports to *CEMA*; the latter would have added $48 million; but the two jointly would have an important interaction term, therefore adding $589 million. Export controls are primarily on high technology items and these are generally only purchased on credit, thus explaining why elimination of restrictions on both exports and credits to finance them increases exports more than the sum of the two actions in isolation. Wolf also estimated that *U.S.* imports from *EE* in 1968 would have gone up from $220 million to $418 million if all *EE* countries were granted *MFN*.

TABLE 4—TOURIST REVENUES (GROSS AND NET) OF EAST EUROPEAN COUNTRIES AND THEIR RELATION TO EXPORTS AND TRADE DEFICITS, 1971

(In Millions of Current Dollars)

Country	Total Tourist Revenue	As % of Total Exports	Total *HC* Tourist Revenue	As % of *HC* Exports	Net *HC* Tourist Revenues	As % of Aver. 1969–71 *HC* Trade Deficit	As % of Aver. 1969–71 Imports from U.S.[a]	As % of 1969–71 Trade Deficit with U.S.
USSR	n.a.	n.a.	96	1.7	72	27	58	114
Bulgaria	113	5.2	41	13.5	36	92	916	2,769
Czechoslovakia	61	1.5	18	2.2	10	10	40	261
Hungary	100	4.0	45	7.4	28	31	91	132
Poland	36	.9	16	1.4	12	—[b]	19	—[b]
Romania	86	4.1	57	7.8	45	26	93	140
Yugoslavia	359	19.8	349	36.9	134	14	89	255

[a] Imports from United States as reported by East Europe, which may differ from *U.S.* reported exports to East Europe.

[b] Trade surplus.

Source: Marer and Tilley.

increase drastically its exports of manufactures would remain in doubt. In addition to the oft-cited problems of quality, modernity, reliability, servicing, packaging, and marketing, they would face a tariff structure that discriminates against labor-intensive manufactured products (Robert Hawkins), and *NTB*s, which are concentrated on agricultural commodities and on those labor-intensive manufactured goods in which the United States and other advanced industrial countries are losing their competitive advantage (Ingo Walter).

Tourism as Source of Foreign Exchange. Several less-developed European countries such as Yugoslavia, Greece, Portugal, and Spain earn hard currency (*HC*) to finance commodity imports through tourism and exports of labor. Though all of our *EE* respondents save the Yugoslavs dismissed the latter option as unrealistic, they unanimously agreed that the former is a real possibility.

Table 4, which summarizes the results from Paul Marer and John Tilley's study of tourism in *EE*, indicates the situation as of 1971, the latest date for which complete data are available. (See also Brown and Marer.) Several findings are worthy of note: (1) only in Yugoslavia were total tourist revenues large relative to total exports; (2) when, however, we concentrate on *HC* tourist revenues, which is a much more important category for our purposes, these revenues become important also in Bulgaria, Hungary, and Romania, accounting for 7 percent or more of exports to hard currency areas; (3) when we compare *net HC* earnings from tourism (total *HC* earnings minus expenditures of the *EE* country's tourists in *HC* areas and the *HC* import content of services provided to *HC* tourists) to the average *HC* trade deficits, we find that net *HC* tourist earnings covered a large proportion of this deficit in all countries, except in Czechoslovakia; (4) if we made the admittedly extreme assumption that all of the net *HC* earnings would be used to finance imports from the United States, then it is clear that a large expansion of these imports would be feasible beyond the relatively low 1971 levels. The assumption is less unreasonable in the case of tourism than it would be for commodity exports, since commodity trade between *EE* and *WE* is frequently tied under formal or informal bilateral agreements, while this is not true for tourism.

Tourist revenues of each of the *EE* countries except Czechoslovakia grew more rapidly than their exports during the 1960's. Equally impressive is the comparison between the ratios of tourist revenues and revenues from commodity exports in *EE* and *WE*. This ratio is higher in all *WE* countries except in West Germany and the United Kingdom than in *EE* and is extremely high in the smaller and/or less-developed countries, like Spain, Greece, Austria, and Portugal (with ratios ranging from 20 percent for Portugal to 70 percent for Spain). Setting aside the political constraints on the expansion of *EE* tourism, it would appear that if *WE* provides any indicator of the potential for the tourist industry, there is considerable room for further expansion of *EE* earnings from tourism.

In view of the past dynamics of *EE* tourism, the comparative statics of *EE* and *WE* tourism, and the ambitious plans to encourage Western tourism in all *EE* countries except Czechoslovakia,[2] it is possible to argue that the tourist industry is one of the more important potential earners of foreign exchange to pay for potential increased *EE* imports from the United States and that its importance relative to commodity exports to *HC* areas is likely to rise in the future.

Credits. Restrictions on credits to finance *U.S.* exports to *EE* represent major barriers to increased *U.S.-EE* trade, a fact confirmed by all our respondents from *EE*, and quantified in most interesting ways in studies by Wolf (1972 and 1973a) and by Josef Brada and Larry Wipf. Machinery and equipment sales, which form the major category of potential *U.S.* exports to *EE*, are virtually always dependent on the availability and cost of credit. Thus, it would seem useful to posit a functional relationship between the availability, i.e., supply, of standard credits at various rates of interest, either constrained by absolute limitations or not, and the volume of *EE* purchases from the United States, and to measure the elasticity of *EE*'s import demand with respect to this credit availability. By way of analogy, income and price elasticities of imports and exports are time-honored concepts in foreign trade theory and empirical work. They play an important role in the analysis of *U.S.* trade with *EE*. "Credit elasticity" is much more difficult to deal with conceptually and especially empirically, but this does not make it less important. We are exploring this concept more fully and will report on some empirical results soon (see Brown, Marer and Neuberger).

Multilateral Exchange. A potential source of foreign exchange to finance the prospective *EE* deficit with the United States is the use of *HC* earned by exporting to other countries. This requires the existence of multilateral trading and clearing patterns. Little systematic work has been done so far on the question of how multilateral is East-West trade, and experts disagree. Franklyn Holzman (p. 684) maintains that East-West trade is largely multilateral, while Samuel Pisar (p. 169) concludes that bilateralism is a permanent feature of East-West trade.

All measures of multilateralism (including the League of Nations, the Michaely, and the Pryor indices), are based on *ex post* analysis of trade data, while multilateralism is an *ex ante* decision to permit multilateral clearing. Thus, a continuous bilateral trade deficit over the years may indicate either a well-established multilateral clearing pattern, or that imports are being financed by invisibles or credits.

According to the standard statistical measures which we applied, East-West trade *appears to be* at least as, if not more, multilateral than trade among the principal *WE* countries. Not being satisfied with the standard approach for the reasons indicated, we have developed a new measure and experimented with some refinements (see Brown, Marer and Neuberger).

---

[2] In 1973 East Germany received an estimated 120 million DM through compulsory currency exchange at the border (*Der Tagesspeige*, Nov. 21, 1973).

REFERENCES

E. **Borissoff** and S. **Sind**, "Projections of U.S. Exports to USSR and Eastern Europe," Res. Note 3, Analysis Division, Bureau of East-West Trade, U.S. Dept. of Commerce, May 1973.

J. **Brada** and L. **Wipf**, "U.S. Trade Controls on Exports to the Soviet Bloc," *So. Econ. J.*, July 1974, *41*, forthcoming.

A. **Brown** and P. **Marer**, "East-West Trade: Old Issues and New Prospects," in Joint Econ. Comm., U.S. Congress, *A Foreign Economic Policy for the 1970's*, Part 6, Washington 1971, 1211–1226.

———, ———, and E. **Neuberger**, "East-West Trade: Options and Prospects," *Reorientation and Commercial Relations of the Economies of Eastern Europe*, Joint Econ. Comm., U.S. Congress, Washington 1974, forthcoming.

R. **Hawkins**, "Tariffs and Tariff Structures: An International Comparison," in Hawkins and Walter, eds., *The United States and International Markets*, Lexington 1972.

F. **Holzman**, "East-West Trade and Investment Policy Issues: Past and Future," in Joint Econ. Comm., U.S. Congress, *Soviet Economic Prospects for the Seventies*, Washington 1973, 660–689.

A. **Malish, Jr.**, "An Analysis of Tariff Discrimination on Soviet and East European Trade," *The ACES Bull.*, Spring 1973, *15*, 43–56.

P. **Marer**, *Soviet and East European Foreign Trade, 1946–1969: Statistical Compendium and Guide*, Bloomington 1972.

——— and J. **Tilley**, *The Tourism Industry in the Soviet Union and Eastern Europe: Travel Flows, Earnings, and Prospects*, Bloomington 1974, forthcoming.

E. **Neuberger**, "Is the USSR Superior to the West As A Market for Primary Products?", *Rev. of Econ. and Statist.*, Aug. 1964, *66*, 287–293.

S. **Pisar**, *Coexistence and Commerce*, New York 1970.

J. **Stankovsky**, *Bestimmungsgründe im Handel Zwischen Ost und West*, Vienna 1972.

I. **Walter**, "Barriers to International Competition" in Hawkins and Walter, eds., *The United States and International Markets*, Lexington 1972.

T. **Wolf**, "The Quantitative Impact of Liberalization of United States Unilateral Restrictions on Trade with the Socialist Countries of Eastern Europe," prepared for the U.S. Dept. of State, XR/RECS-3, Feb. 16, 1972.

———, "A Note on the Restrictive Effect of Unilateral United States Export Controls," *J. Polit. Economy*, Jan.–Feb. 1973a, *81*, 219–222.

———, "Effects of U.S. Granting of Most Favored Nation Treatment to Imports from Eastern Europe: The Polish Experience," *ACES Bull.*, Spring 1973b, *15*, 23–42.

Joint Econ. Comm., U.S. Congress, *Reorientation and Commercial Relations of the Economies of Eastern Europe*, Washington 1974, forthcoming.

PERSPECTIVES ON TWENTIETH CENTURY ECONOMIC THEORY*

Capital Controversies: Ancient and Modern

By JOHN HICKS**

In what must by the rules be a fairly short paper, I cannot make a survey of all controversies about capital, from (say) Ricardo-Malthus to Joan Robinson-Solow. All I can attempt is something much more modest. What I propose to do is to take one particular point, which has figured (as I shall show) in many such controversies, and to use it as a means of pulling a part of the story together. It is a matter which would seem to be appropriate for discussion at this joint meeting, for it is an interesting illustration of the way in which the history of his subject can be of use to the modern economist.

Economists do indeed have a special use for the history of economics, something more than the general use that can be made of their own history by students of other subjects such as mathematics and the natural sciences. The history of science, certainly, is no mere antiquarianism; one is learning science when one learns in what ways scientific discoveries have been made. The history of economics has that use, and it has other uses. It has, of course, a pure historical use; the greatest economists, Smith or Marx or Keynes, have changed the course of history; they are as worthy the attention of the pure historian as Louis Napoleon or Woodrow Wilson. But this again is not the economist's use. That is something different.

Economics is a social science, and a particular kind of social science, in that it is concerned with the rational actions, the calculated actions, of human beings, and with their consequences. This has the result that those whom we study can hear what we say. We may speak to each other in our private languages, but private conversations are no more than goods in process; while we speak only to each other we have not finished our job. The ideas of economics, the powerful ideas of economics, come from the market-place, the "real world," and to the "real world" they go back. So there is a dialogue between economists and their subject-matter. It is a dialogue in which there are important intermediaries; statisticians are one kind of intermediary, journalists another, accountants (as we shall see) another; the economist-statistician and the economist-journalist do much of the intermediation themselves. In the course of the dialogue ideas acquire associations; they cease to be free ideas, which can be defined at choice. It is not in our power to say with Humpty-Dumpty, "When I use a word it means just what I choose it to mean"; we cannot escape the associations.

I do not mean that there is not such dialogue, and such associations, in the case of other social sciences. Clearly there is; in political science, say, as much, if not more,

---

* The paper given at this session by Joseph J. Spengler of Duke University will appear in the *Journal of Economic Issues*.

** All Souls College, Oxford.

than in economics. Political ideas are indeed so rich in associations that the study of politics seems sometimes to consist of little else. Economics has (relatively) much more that is positive to offer, but we should not allow our passion for quantification to blind us to the fact that economic ideas share this characteristic of political ideas and do so for the same reason.

We cannot escape the associations, but we can try to understand them, so as to be masters of them. That is what, in my view, the history of economics is for. It is what it is for, for the economist. We need to know the history of our concepts in order to know what it is that we are handling.

The history of economics, so understood, cannot be discovered by poring over old textbooks, even old "classics." That is no more than a part of what has to be done. The books must be read against their background, the events which prompted the analysis, and what happened to the analysis when it went out into the world. All that is part of the tradition which we have inherited, and from which, if we are to do our job, we cannot escape.

I turn, I hope in that spirit, to my particular topic.

One must begin with a distinction which has been fully understood only in quite recent years. Suppose we start by saying (as many would say) that the capital of an economy is its stock of real goods, with power of producing further goods (or utilities) in the future—the stock of such goods existing in the economy at a moment in time. That, in strictness, is no more than a list—what in English, but not in American, we would call an inventory. How do we aggregate it, as for macroeconomic purposes we have to do? We cannot aggregate except by adding money values; how do we deflate that money aggregate, so as to get a measure of Real Capital? There are, in principle, two main ways.

The first, which on the analogy of other aggregates, such as consumption, may seem the more natural, is to deflate by an index of the prices of the capital goods themselves. (It may not be easy to find a suitable index, but that is not here my concern.) It is, theoretically, a possible measure; to distinguish it from the other, to which I shall be coming, I call it Volume of Capital. It has the property, it will be noticed, that as between two economies which have capital stocks that are physically identical, Volume of Capital must be the same.

It may well be objected, and has often been objected, that Volume of Capital misses the essential fact about capital—that the utilities of capital goods are indirect. The values of capital goods are derived values, capitalized values of future net products. If these future products are valued at current prices of products, the resultant capital values should be better indicators of the true values of the capital goods than their actual market values, which are influenced by expectations of changing values of products. These "corrected" capital values could be aggregated; but since they have been built up from product prices, not capital good prices, they can only be deflated by an index of product prices, if they are to be used as a measure of Real Capital. Real Capital, in this sense, does not have the invariance property; it may be changed, without any change in the physical goods, by the mere admission of new information. I call this other measure Value of Capital.

When the distinction is expressed in this statistical (or quasi-statistical) manner, it would seem that we need have no difficulty in living with it. We can keep both in play, using one for some purposes, the other for others. We may nevertheless have some difficulty in explaining ourselves. Both, as described, are measures of Real Capital; but what is Real Capital? We cannot say that the two measures are two measures of the same thing—as one may say that one buys a pound of something, or a dollar's worth of it, when the price of the thing is a dollar a pound. The Real Capital that is being measured is different.

If it is capital in the volume sense that is being measured, capital is physical goods; but in the value sense capital is not physical goods. It is a sum of values which may conveniently be described as a Fund. A Fund that may be embodied in physical goods in different ways. There are these two senses of Real Capital which need to be distinguished.

I do of course borrow the term Fund from the history, and to the history I now turn. I am going to maintain that the distinction is quite ancient; it divides economists, ancient and modern, into two camps. There are some for whom Real Capital is a Fund—I shall call them Fundists; and there are some for whom it consists of physical goods. It is tempting to call the latter Realists;[1] but since one wants to emphasize that both concepts are *real*, this is not satisfactory. I shall venture in this paper to call them Materialists. (Materialists, I mean in the sense of Dr. Johnson's refutation of Berkeley's idealism—"striking his foot with mighty force against a large stone, till he rebounded from it, *I refute it thus.*" There will be some at least of my Materialists who are worthy followers of Dr. Johnson.)

One of these was Edwin Cannan, the teacher of Lionel Robbins and the founder of the economics school of the London School of Economics and Political Science. A beautiful illustration of the opposition with which I am here concerned is to be found in Cannan's comments on capital in Adam Smith.[2] Cannan was convinced that Smith was in a muddle. I do not think that he was in a muddle; he was simply a thoroughgoing Fundist, and to the Materialist Cannan the Fundist position was quite incomprehensible. Cannan: "The capital [in Smith] is often spoken of as if it were something other than the goods themselves." That is just the point.

Not only Adam Smith, but all (or nearly all) of the British Classical Economists were Fundists; so was Marx (how else should he have invented "Capitalism"?); so was Jevons. It was after 1870 that there was a Materialist Revolution. It is not the same as the Marginalist Revolution; for some of the Marginalists, such as Jevons and Böhm-Bawerk, kept the Fundist flag flying. But most economists, in England and in America, went Materialist.[3] Materialism, indeed, is characteristic of what is nowadays reckoned to be the "neoclassical" position. Not only Cannan, but Marshall and Pigou, and J. B. Clark, were clearly materialists. Anyone, indeed, who uses a Production Function, in which Product is shown as a function of labor, capital and technology, supposed separable, confesses himself to be (at least while he is using it) a Materialist.

What about Keynes? Keynes, of course, was brought up as a Materialist, and there are no more than slight signs, in the *General Theory*, that he had departed from the Materialist position. So it is perfectly possible to be a Keynesian and yet to be a Materialist. But the rethinking of capital theory and of growth theory, which followed from Keynes, and from Harrod on Keynes, led to a revival of Fundism. If the Production Function is a hallmark of Materialism, the capital-output ratio is a hallmark of modern Fundism. That, in the briefest of outline, is the story; how it happened I shall now attempt to explain.

---

[1] As I have done myself in a short passage in my *Capital and Time*, p. 13. I have since become convinced that Materialist is better.

[2] Cannan, pp. 145–50.

[3] But there was at least one important American Fundist, F. W. Taussig. Irving Fisher is harder to place, since he, at least sometimes, could see both sides. But it is interesting to find that Cannan thought Fisher, like himself, to be a Materialist.

Let us go back to the Classics. Why were the Classical Economists Fundists? It is not easy to see, just from looking at their works; they take their Fundism so much for granted that they do not need to justify it. Surely the reason is that it came to them from outside—from business practice, from accounting practice.

Even to this day, accountants are Fundists. It is not true, accountants will insist, that the plant and machinery of a firm are *capital;* they are not capital, they are assets. Capital, to the accountant, appears on the liabilities side of the balance sheet; plant and machinery appear on the assets side. Capital, accordingly, is a Fund that is embodied in the assets.

The origin of accounting is in the business of the merchant; accounting categories, to this day, bear the mark of their mercantile origin. It was the merchant who was the original Fundist. It is the merchant who thinks of his capital as a Fund that is invested in a stock of salable goods. It is in the Fund sense that capital "circulates"; the physical goods do not circulate, but the Fund does. It is the Fund that is "turned over." The stock of goods in the merchant's possession is one thing (the most he will admit is that it is the form that is taken by his capital at the moment); his capital, he will surely say, is something more permanent.

These were the business terms which came naturally to the Classical Economists. They had no reason to depart unnecessarily from the businessman's language. This was the system of thought that the businessmen of their time were using; they just followed it. It is true that they were thinking of the whole (national) economy, not of the single business; but this, they surely felt, made no difference. It did not need to make much difference. They could think (as Henry Thornton, in particular, most surely did think) of the whole economy as having a balance sheet, constructed by consolidation of the balance sheets of the single businesses; in the consolidation, the liabilities of one unit would be cancelled against the assets of another, but no item would be transferred from the liabilities to the assets side. So, even when all debts and paper claims had been cancelled, there would remain on the assets side the real goods (and balance of external claims), on the liabilities side the Capital—still a Fund. It need not be thought of as a debt owed by the nation to itself; it is the same kind of thing as the Capital of the single business.

The way would thus be open for Classical Fundism if the whole economy consisted of merchants; how far, however, could it cope with businesses of other kinds? It was necessary, from the start, to deal with businesses of other kinds; but for the first of the extensions that were needed, Fundism did not do at all badly. It is often thought that the notion of capital as "advances to laborers" took its origin from observation of agriculture, so it is labelled physiocratic; and it is true that if one looks only at the economic literature it is with the French physiocrats that it seems to come in. But so far at least as the British Classical Economists are concerned, it is more convincingly interpreted as a fitting of agricultural experience into the mercantile pattern. The farmer, like the merchant, "turns over" his capital, buying the services of labor, as the merchant buys his stock in trade; selling the product of that labor when it is ready to be marketed. So the Fundist concept of capital could be carried over to agriculture, surviving the transition. It seemed, on the whole, quite a good fit.

The farmer, of course, used land as well as capital, but that land was a separate factor of production no one doubted. The rent of the land must be deducted from the "gross" profit on any agricultural operation before the "net" profit was arrived at; it would be the rate of net profit on the Capital Fund that competition would tend to equalize.

Classical economics was three-factor economics, and we can now see that the triad had deeper roots than is commonly supposed. Labor is a flow, land is a stock (as stock and flow are used in modern economics); but capital is neither stock nor flow—it is a Fund. Each of the three factors has its own attribute, applicable to itself but to neither of the others. Labor works *on* land *through* capital, not on capital nor with capital. The place of each of the factors in the productive process is sharply distinguished.

The Classical Economists, so interpreted, are rather consistent; among their successors, in the latter part of the century, consistency disappears. Not all of those who went Marginalist went Materialist, and those who did go Materialist did so in different ways. The case of Walras, for instance, is quite peculiar. I feel fairly sure that he is to be reckoned as a Materialist; but the reason for his Materialism is his interest in particular capital goods, appearing (of course) in his work as a part of his general determination to work with an *n*-goods model. Yet his Materialism may antedate that determination, and may have been one of the things which impelled him towards it. He says that he took his view of capital from his father and Auguste Walras (writing in 1849) would certainly seem to have been a Materialist, even an extreme Materialist. "Capitals" (*capitaux*) to him are capital goods; "incomes" (*revenus*) are income goods; they are distinguished by multiple (*successive*) uses against single uses and by that alone. Carriages, carts, steam-engines are *capitals;* a glass of wine, a round of beef, a candle are *incomes.* "*Le revenu, ainsi que son nom l'indique, c'est ce qui revient, or, ce qui revient, c'est ce qui s'en va.*" (pp. 53–54).[4]

One must yet beware, in father and son alike, of mistaking for a theoretical approach what is no more than a peculiarity of the French language, the restrictiveness, the deliberate restrictiveness, of its vocabulary. It may well be that much French Materialism is only apparent, a matter of linguistics rather than economics.

The case of Marshall is here more interesting. Marshall's Materialism is much more clearly to be explained by events—by the now achieved Industrial Revolution—the rise in importance of plant and machinery. The Classical schema, as we have seen, began with trade and was extended to agriculture; so long as stocks and work in progress were the main part of the manufacturer's physical assets it could be extended, in much the same way, to manufacture also. But when a large part of his capital became *fixed* in plant and machinery, a candidate had appeared for factor status, which did not fit into the classical triad. What was to be done?

Before considering what happened in the economics, it will be useful to turn again to the accounting aspect. The rise of the Machine had already presented the accountant with a parallel problem. What did the accountant do about it?

So long as he had nothing to consider but mercantile transactions, his task in principle was simple. For it is characteristic of the business of the merchant that it is divisible into separate units. Every bale of cotton or pound of cheese which ever forms part of his stock is acquired at a particular date and sold at a particular date; purchase, retention and sale constitute a separable transaction. (Complete separation is of course not attained in practice, since there are overheads which have to be allocated; but it is so nearly attained that it sets the pattern.) So the accounts of the merchant may be regarded as a bundle of separate accounts. Purchases and sales are indeed going on continually, so that if they are set out in a time sequence, the separate accounts will overlap; it is only if an account were prepared for the whole history of the business, from first setting-up to final closing down, that a record of purchases and sales would tell the whole story. In any arbitrary, say, annual, period there must be transactions which have started before the beginning of the period, but are completed within the period; there must similarly be transactions which are begun within the period but not completed when it is ended. These, however, in the case of a mercantile business cause the minimum of trouble. They can be dealt with by the accepted rule of never taking a profit on any transaction before it is completed. The initial stock of the year will then be brought in *at cost*; and the final stock will be valued at cost in the same manner.[5]

---

[4] ("Revenue [the return], as its name indicates, is that which comes back; now, that which comes back, is that which went off."—Editors.)

But what was to be done, on these principles, with plant and machinery? The use of land, being regarded as a permanency, could be brought in as a regular charge; but the plant and machinery is not expected to last indefinitely, though its use is spread over a time which is longer than the accounting period. It is important to observe that it is the extension of the use to a duration which is longer than the accounting period which creates the difficulty. There would be no difficulty, here as in the mercantile case, if the account were drawn up for the whole life of the business, from first setting-up to final closing-down.[6] It is for the annual account that there is the problem. The cost of the machine has to be set against a series of sales, the sales of the outputs to which it contributes, but some of these sales are sales of the present year, some are later and some, maybe, earlier. There is thus a problem of imputation; how much is to be reckoned into the costs of this year, and how much into the costs of other years? It is just the same problem as the allocation of overheads, and to that, as is now well known, there is no firm *economic* solution.

Neither has the accountant found a solution—only a name and a set of (essentially arbitrary) rules. The "depreciation quotas" must add to unity, but that is all that is known, at all firmly, about them. The form of the account is preserved, but only by bringing in, as the capital which is supposed to be invested in the machine at the beginning of the year, that part which has not been absorbed (by being allocated as a cost to the output of preceding years) and by reckoning as the capital invested at the end of the year that part which has not been absorbed in those years nor in this year—so that it is left to be carried forward to the future. This is in fact what accountants did (probably what they had to do) as soon as they were confronted with the problem. It is what they still do, even to this day.

We have had plenty of opportunity in the present century to understand how arbitrary the accountant's depreciation quotas are. We have seen them battered by inflation, and we have seen them manipulated by tax authorities in the interests of fiscal policy. Late nineteenth century economists had much less of this experience, so it is not surprising to find that they began by taking the accountant's depreciation quotas much too seriously. That is true of both schools. It is true of the surviving Fundists (such as the Austrians) who conceived of investment in fixed capital as equivalent to a bundle of investments in circulating capital. If one-tenth of the cost of a machine of ten-year life could be imputed as cost to the product of each year, the machine was equivalent to ten investments in circulating capital, one with an investment period of one year, one of two years . . . and one of ten years. It was the accountant's depreciation quotas which did the trick.

---

[5] There is of course the qualification that an expected *loss* may be taken in advance: "cost or market value whichever is the lower." But this does not affect the principle.

[6] It was the realization that the economist, unlike the accountant, need not be bound down by annuality, unless he chooses, that prompted me to write *Capital and Time.*

It is also true of Marshall. He also relied upon the accountant's "solution," but in a different way. It was an essential element in his concept of long-period equilibrium. In the short period, Marshall tells us, when "the producers have to adjust their supply to the demand as best they can with the appliances at their disposal," the "income" derived from those "appliances" is a *quasi-rent*.[7] It is called a quasi-rent by analogy with the rent of land: in all of his short-period theory Marshall has Ricardo's rent theory very much in mind. Now the rent which is determined as a surplus, in Ricardo's manner, makes no allowance for depreciation; in Ricardo, land being "indestructible," no such allowance is of course required. Marshall, however, does think of a deduction being made, *even in the short period*, though Ricardo has given him no help in determining what that deduction should be. It is true that Marshall is so reticent on the matter that he can easily be misunderstood; one has to read him quite carefully to discover whether quasi-rents are to be taken gross or net. He does however say in a definitional chapter of *Principles*:

> We cannot properly speak of the interest yielded by a machine. If we use the term *interest* at all, it must be in relation not to the machine itself but to its money value. For instance if the work done by a machine which cost £100 is worth £4 a year net, that machine is yielding a quasi-rent of £4 which is equivalent to interest at four per cent, on its original cost; but if the machine is worth only £80 now, it is yielding five per cent on its present value. [pp. 74–75]

This seems conclusive.

It is indeed quite remarkable how little there is in Marshall's book about depreciation. There is a footnote in which he recognizes that it is a problem, but the footnote just ends with a reference to an accounting textbook.[8] He has evidently decided that for his purposes, the accountant's solution will do. Gross can in that way be reduced to net, and it is net returns that are equalized by competition in the long period.

That is what Marshall says, but in the forty years which followed on the publication of his *Principles*, the strangest things must have happened to teaching on this point, even among his closest followers. One had heard rumors that there was a good deal of confusion between "gross" and "net" among Cambridge teachers, and a strong piece of evidence in support of them has now come to hand. In his chapter on the marginal efficiency of capital in the *General Theory*, Keynes is careful not to call his Q's quasi-rents, and rightly so, since they are gross of depreciation, so they are *not* what Marshall meant by *quasi-rent*. But why are they symbolized by Q? It has naturally been assumed that Q stands for quasi-rent; so the Q's are called quasi-rents in many post-Keynesian writings. We now know that in an earlier draft the Q's *were* called quasi-rents.[9] Keynes made a correction, or semicorrection, in the final version, but the confusion is betrayed.

---

[7] Marshall, p. 376.

[8] Marshall, pp. 354–55.

[9] Keynes, pp. 425–26.

The fact is that until the new wind began to blow, in the mid-twenties, very little interest had been taken in Cambridge in capital theory. Take the case of Pigou. One of the remarkable changes introduced into the later editions of *The Economics of Welfare* is a chapter entitled "Maintaining Capital Intact." There was no such chapter in the first or second edition. It makes its appearance in the third (1928), by which time the problem of defining saving, in conditions of changing prices, had been brought to light in the work of Robertson (1926). In the fourth edition (1932) it is considerably altered; and there is a further version, in an important article (1935) which looks as if it was intended as the basis of another revision.[10] Pigou, it is clear, was very bothered; and one can see why.

It was only a part of the problem of capital with which he thought himself to be concerned; he was trying to look at that one part in isolation. In spite of its "Welfare" coloring, the subject of his book was the Social Product (or "dividend" as he calls it): how it is measured, what makes it large or small, how it is distributed. The Social Product of one particular period ("year") is considered, almost, in isolation. Having chosen that way of posing his problem, he is led (almost inevitably) to what would nowadays be called a "Production Function" approach. Even so, he might have found himself confronting the general problem of measuring capital—for how can we make a static comparison, between two economies whose capital stocks are different, without having some means of comparing their capital stocks? Pigou did not, at first, raise this wide question, though Robertson, at much the same date, could already see that it was involved.[11] He confined himself to what arose in the measurement of income, in the single year—the measurement of the investment component, the reduction of gross investment to net. (This, of course, already implies a problem of capital measurement—the comparison of the beginning-year and the end-year capital stocks.)

Pigou's approach is strictly Materialist. He does indeed recognize that the business concept of capital is different, but

> for economics the stock of capital existing at any time is a collection of objects, the extent of which is a purely physical fact. . . . the size of the stock is not . . . *affected* by its value; it is exactly the same . . . whether that value is large or small. [1935, p. 235]

He goes on to draw from this the important conclusion, echoes of which are to be found in many more recent writings:

> A distinction should be drawn between changes which, while leaving the element still as productive as ever, bring nearer the day of sudden and final breakdown, and physical changes which reduce its current productivity and so rentable value. With the former sort of change, until the breakdown occurs, the capital stock is, I suggest, best regarded as intact, just as it is best regarded as intact despite the nearer approach of a day that will make a part of it obsolete. [1935, p. 238]

I shall not discuss Pigou's treatment in detail; enough has been said to indicate its Materialist character. For the purpose which Pigou had in mind, it may well be defensible (as I shall show), but it can hardly be regarded as general; so it is not surprising that there were quarters where it was not well received. I think not so much of the "user cost" chapter in Keynes's *General Theory*, which I regard as an unsuccessful (and for Keynes's purpose unnecessary) attempt to bridge the gap. The most direct statement of an opposing view came from Professor Hayek.

---

[10] Pigou (1935).

[11] See Robertson's paper "Wage Grumbles," reprinted in Robertson (1931). I have a particular affection for that paper since it was through it that I myself first came into touch with its author.

Hayek's first paper on the subject was already in preparation when Pigou's article appeared in 1935.[12] He further developed it in his *Pure Theory of Capital* and in a paper in *Economica* which is a reply to a further note by Pigou.[13] Hayek, of course, was a Fundist, but a very sophisticated Fundist, deeply preoccupied with the problems of ignorance and uncertainty which come to the fore as soon as one thinks of capital value as being determined by expectations of the future. It was the omission of this aspect which set him against the Materialism of Pigou. His striking illustration—of the machinery installed to produce a fashion good, required within the year but not afterwards, so that at the end of the year the machine which is physically intact has lost its value—is directed to show the economic irrelevance of physical identity. *In general*, so it seems to me, Hayek is right.[14]

I shall not attempt to carry the story further, except to suggest that we may get a useful light upon more recent differences if we look at them in terms of this interesting controversy between Pigou and Hayek. Pigou, I think one can now see, was working within a model—a rather restrictive model but a useful model, a model which for certain purposes we would most of us still wish to employ. Its central concept was the Social Product, a *measurable* Social Product. It has now become clear that one cannot measure a Social Product of the Welfare type with which Pigou was concerned unless one assumes that wants are unchanging—or that some kind of Social Welfare Function is unchanging—and that cuts out the sort of variation over time which is the root of Hayek's example. In Pigou's world the Hayek problem could not arise. In general, of course, it does arise. There are bound to be odd cases which will not fit into the Pigou model.

I do not mean to imply that attention to odd cases is the principal reason for the revival of Classical Fundism. It is one of the reasons; there are plenty of economists who make their living by trading in odd cases, which make splendid subjects for little notes in journals. There are other reasons which are more substantial. Fundism, as I have emphasized, is the businessman's concept of capital; social accounting, which has brought economists into closer relation with business accounting, was bound to induce a revival of Fundism. So was attention to planning. Planning is forward-looking, and the Fundist concept of capital is forward-looking; they fit in. For such and such a development, how much capital will be required? It is capital in the Fund sense to which such questions are relevant.

When Fundism is looked at in this light, we can see that there may be a place for Materialism also. If Fundism is forward-looking, Materialism is backward-looking, for it is concerned with the capital goods existing at the moment, goods which have been brought into being in the past. "Bygones are bygones," but there are still some purposes for which we have to go to the past. Our statistics, in particular, always relate to the past. One cannot measure capital, statistically, excepting in terms of its history, valuing, of course, not at historical cost, but at doctored historical cost, or replacement cost. There is a striking example of this in the book by J. R. S. Revell, *The Wealth of the Nation* (1967), an inquiry conducted in Cambridge, England, a place that must often appear on this side of the Atlantic as a headquarters of the New Fundism. Revell's calculation of the National Capital of Britain would have appealed to Pigou;[15] he had to be Materialist, because as a statistician he could be nothing else.[16] There is nothing else that can be used in macroeconometrics; so it is macroeconometrics itself which is on trial —but that, perhaps, is as it should be.

---

[12] Hayek (1935).

[13] See Hayek (1941a) and Pigou (1941).

[14] In my own contribution to the same discussion (Hicks, 1942), I came down mainly on Hayek's side. The positive part of my paper is reprinted in *Capital and Time*, pp. 164–66.

REFERENCES

E. Cannan, *A Review of Economic Theory*, London 1929.

F. Hayek, "The Maintenance of Capital," *Economica*, Aug. 1935, *2*, 241–74.

———, "Maintaining Capital Intact," *Economica*, Aug. 1941a, *8*, 276–80.

———, *Pure Theory of Capital*, London 1941b.

J. R. Hicks, "Maintaining Capital Intact," *Economica*, May 1942, *9*, 174–79.

———, *Social Framework*, 4th ed., New York 1971.

———, *Capital and Time*, Oxford 1973.

J. M. Keynes, *Collected Writings*, London and New York 1971–, *13*.

A. Marshall, *Principles*, 8th ed., London, 1922.

A. C. Pigou, *Economics of Welfare*, 3rd and 4th ed., London 1928, 1932.

———, "Net Income and Capital Depletion," *Econ. J.*, June 1935, *178*.

———, "Maintaining Capital Intact," *Economica*, Aug. 1941, *8*, 271–75.

D. H. Robertson, *Banking Policy and the Price Level*, New York 1926.

———, *Economic Fragments*, London 1931.

F. W. Taussig, *Wages and Capital*, New York 1896.

A. Walras, *Theorie de la Richesse Sociale*, Paris 1849.

---

[15] I am aware that in his last note on the subject, Pigou (1941) dissociated himself from the replacement cost measure of capital, pointing to the case where the article to be replaced "had become impossible to produce" as a reason for rejecting it. I do not think he was well advised to make this concession. He would have done better to insist that it was one of the "odd cases" that he was leaving out.

[16] For further discussion of this matter of measurement, see Hicks (1973), Chapter 13, and (1971), Appendix D.

IS ECONOMIC THEORY WITH IT?

On the Relevance of the New Economics of the Family

By GARY S. BECKER*

My answer to the question posed by the title of this session is "yes" both in the sense that economic theory is very relevant in understanding and predicting significant aspects of human behavior, and that it is used for this purpose by a large proportion of economists. This is not to deny that for some economists economic theory is a game played for its own sake, nor that the economic theory of the year 2000 will be decidedly better than the theory of 1970. Both are probably inevitable in any discipline centered in academic institutions. But the game mentality is not the dominant one, thanks in large part to the increasing interaction between theoretical and empirical analysis which has developed in academic institutions. For the same reason, economic theory today is far more "with it" than the theory of 1900 was.

One of the most significant recent developments in economic theory has been its application to behavior outside the monetary market framework: for example, decisions about family size, the labor force participation of married women, illegal behavior, political behavior, statistical inference, or the accumulation of education and human capital (which has been used to explain race and sex differences in income). This extension of the scope of economic theory resulted partly because a number of economists—myself included —were dissatisfied with the state of theory as of twenty years ago. I believe that this extension has not only increased the theory's relevance, but also strengthened its structure because of the need to adapt the theory to novel situations.

Let me illustrate these developments by discussing briefly one application of them, which may go a small way toward convincing the skeptics about the value of economic theory. (More complete documentation of what I say can be found elsewhere.)[1] This application, which deals with the behavior of families, is chosen partly because recent studies based on it have been so successful, and partly because a distinguished past president of the Association, John Kenneth Galbraith, has recently criticized economic theory for neglecting the behavior of families.[2]

The so-called new economics of the family emphasizes that the time of different family members is their primary scarce resource and replaces the assumption that families are passive consumers of what they purchase with an assumption that families both produce and consume objects of choice, called "commodities," using inputs of their own time and purchased goods and services. Each family is assumed to maximize utility by producing and consuming different commodities, subject to its production functions, the limited time of each member, their wage-generating functions, and other variables.

---

* University Professor of Economics, University of Chicago.

[1] For example, see Jacob Mincer, Becker, and T. W. Schultz (1973 and 1974). For a survey contrasting the old and new economics of the family, see Robert Michael and Becker.

[2] See Galbraith. Robert Michael of the National Bureau of Economic Research called my attention to this article and has written a comment on it that is available to the interested reader.

A number of significant implications follow from this approach. For example, a rise in the wage rate of one family member relative to another would, through the usual substitution effects, induce a family to use less of the first member's time in home production and more of the second member's time. Put differently, more of the first member's and less of the second member's time would be supplied to the labor force. This reasoning accounts for the fact that a married woman is more likely to be in the labor force either when her own wage rate is higher or when her husband's wage rate is lower. This reasoning also explains the less well-known finding that husbands work fewer hours when their wives' wage rates are higher. These findings in turn go a long way toward explaining why married black women work more than married white women at the same time that married black men work less than married white men: wage rates of black women are much closer to those of white women than wage rates of black men are to those of white men.

Children are obviously not purchased in the market place, but are "produced" by each family through inputs of various market goods and services and the time of family members. The value of the time of parents, especially of mothers, is a major cost of having and rearing children. Therefore, given the basic theorem of demand theory about negatively inclined demand curves, it is not surprising that the number of children a family has is strongly negatively related to the mother's value of time, as measured by her wage if she is in the labor force (see T. W. Schultz, 1973 and 1974). Much of the secular decline in average family size in the United States is a consequence of the secular rise in the value of the time of married women.

In general, married women spend more time at nonmarket activities, especially at child care, than married men do, which is the counterpart of married women spending less time in the labor force. My earlier discussion of labor force participation explains this difference by arguing that married women substitute nonmarket for market time because their wage rates are lower than those of married men. However, recent work on the accumulation of human capital (see, e.g., Jacob Mincer and Solomon Polachek) indicates both that most of the difference in wage rates between married men and women results from less accumulation of marketable human capital by married women, and that women accumulate less partly because they spend less time in the labor force. (Presumably women accumulate more human capital that is useful in the nonmarket sector than men do because women spend more time there.)

The conclusion is, therefore, both that married women spend less time in the labor force partly because their wage rates are lower, and that their market wage rates are lower partly because they spend less time there! One is naturally led further to ask, if married women worked as much as married men, would the wage rates of married men and women be about the same? If so, how is the sexual division of labor between the market and nonmarket sectors explained?

One answer, gaining increasingly in popularity, is that married women work less because the market discriminates against them and because women have been "brainwashed" against market participation. An alternative answer is that even if women are not intrinsically less productive in the market sector, they may be intrinsically more productive in the nonmarket sector, especially with respect to child care. Of course, even this alternative view would admit that a sharp sexual division of labor has become less necessary over time as average family size has declined and as industrialization has changed the nature of work. Moreover, perhaps discrimination and "brainwashing" have made the responses to these changes—e.g., increased labor force participation of married women, and higher divorce rates—slower and less complete than they should be. It is not possible yet to determine the correct explanation of the sexual division of labor between married men and women, but the power and relevance of the new economics of the family lead me to expect that the answers to this basic question will be forthcoming before "very" long.

Instead of preaching about the relevance of economic theory, I have spent my limited time on a specific example, the new economics of the family. I could equally well have chosen many other examples: better-known ones within the class of applications to behavior outside the monetary market sector, like the economic analysis of education, or less well-known but highly promising ones within this class, like the economic analysis of government regulation and politics.

I am not trying to suggest that everything is for the best with economic theory. Indeed, I could spend a long time elaborating my views about what is wrong. But a field should be judged more by its successes than by its failures. And economic theory is a powerful tool for understanding human behavior, especially when tied to empirical work and oriented (if only indirectly) toward relevant social issues.

REFERENCES

G. S. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*.

J. K. Galbraith, "The Economics of the American Housewife," *The Atlantic*, Aug. 1973.

R. T. Michael and G. S. Becker, "On the New Theory of Consumer Behavior," *Swedish J. of Econ.*, Dec. 1973, *75*.

J. Mincer, "Labor Force Participation of Married Women," *Aspects of Labor Economics*, Natl. Bur. of Econ. Res., Princeton 1962.

——— **and S. Polachek**, "Family Investments in Human Capital: Earnings of Women," *J. Polit. Econ.*, Mar.–Apr. 1974, *82*.

T. W. Schultz, ed., "New Economic Approaches to Fertility," *J. of Polit. Econ.*, Mar.–Apr. 1973, *81*, Supp.

———, ed., "Marriage, Family, Human Capital, and Fertility," *J. of Polit. Econ.*, Mar.–Apr. 1974, *82*, Supp.

Economic Theory and Its Discontents

By VERNON L. SMITH*

The critics of economic theory charge that, while theorists fiddle, the empire burns. We are urged to abandon what we have considered to be economic science and rise to an attack upon the agonizing social problems of the day. Ultimately, I believe, there is only one way to change the course of economic analysis, and that is for its disheartened critics, or others, to chart and pursue a new direction. If it leads to something insightful, it will command a respectable following. But generally the critics have not chosen to change the course of economics in this manner. Instead, they have paid us the high compliment of attacking what we have done and urging us to turn our efforts to the ends which they seek. Unfortunately, most of us have remained stonily silent about what they have said.

Although it may be uncomfortable for some, I think it is important to listen carefully to these critics. Amidst the fulminations of emotional rhetoric there is perhaps a suggestion of deeper issues. If there are indeed scientific issues contained in the bluster, they should command professional attention. I am particularly convinced of this after reading the *Proceedings* of the year before last which—apart from some improvements over previous such efforts—do not inspire confidence that the messengers of alarm can build a theoretical foundation in any direction. They are good at expounding alleged faults, but bad, on the whole, at correcting them; good at suggesting new directions, but bad at charting them. If there is a new economics in the future (and I predict that there is) it will not be born of sloppy new theory, any more than it will be born of old textbook theory, although such a poor start could well motivate new and more interesting efforts. I hope that it does. But neither will new and novel economics be born of an attack upon the scientific ideals of economic theory. It may indeed be true that those of us who try hardest to maintain scientific objectivity will sometimes find that our product is more infected by our values than we would wish. But that hardly provides license for us all to start raking the same muck together. The process of professional evaluation of scientific results, refereeing, e.g., is the only means by which the relationship between values and research can be disciplined. Yet, oddly, it is those who ridicule the ideal of a value-free core of economic methodology who are the most cynical about this process. Such processes are many times imperfect and frustrating, particularly to the innovator, but they are the bedrock of community. Had men never developed the sensitivities that support scientific ideals and professional processes, I fear that we would still be eating each other and justifying it on grounds of relevance.

Our business is the study of mechanisms and institutions for the creation as well as control of individual and social opportunities. This is what markets, voting, courts, legislatures, committees, clubs, associations, and contractual agreements are all about. Man's propensity to truck, barter, and exchange is but a special case of his propensity for social exchange in nonmarket and institutionally constrained market contexts. Perhaps the signal in the critics' outrage is that economic relevance

---

* California Institute of Technology and University of Massachusetts.

can be expanded if we examine such cultural problems as racial discrimination, female servitude, poverty, and pollution. But to suggest that economic analysis will be found relevant to these problems is not to assert that economics can necessarily be used to cure such social ills. Sometimes one of the important functions of dispassionate economic analysis is to make clear that certain social problems may not have a solution. Economics is relevant to the study of inflation and unemployment, but these social problems would seem to be utterly insolvable short of total regimentation if society insists that their "solution" must mean zero percent rates for each. Gary Becker has shown that economics may be relevant to the institution of marriage, but, if divorce is thought to be a social problem, I doubt that it will be solved by use of his theory of marriage. I happen to believe that economics is relevant to the great animal extinctions of ten to fifteen thousand years ago during the age of Paleolithic big game hunting, but the theory of common property resources by itself is unlikely to solve the problem of species extinction in, say, international waters. In fact the theory makes plain the enormous poaching incentives that undermine repeated efforts to regulate the harvesting of prey.

Some of the great social problems about which the radicals are bouncing off the walls have scientific content which in stylized form almost certainly will be found to be capable of economic analysis. To the extent that such problems are rooted in the way institutions have been designed, the right kind of analysis may suggest new and more flexible organizations, but to the extent that such problems are rooted in the human condition, cures may not be forthcoming.

I believe that the microeconomic theory of the pre-1960's is a dead end. I would describe that theory as the economics of individual and social choice in which it was assumed implicitly that certain conveniently chosen activities had infinite information and/or transactions cost, while certain other conveniently chosen activities had zero information and/or transactions cost. Where these costs were infinite, we named the associated activities "externalities," and we cringed or naively called for government intervention; where these costs were zero, we named the activities "markets," and we said "God Bless." That paradigm of thought goes a long way toward explaining why so many good theorists think with their heads when teaching or writing about theory, but with their guts when talking about policy. The paradigm had to end, if not through vision, then through the utter boredom caused by continual work on the same problem in different disguises. The new microtheory will, and should, deal with the economic foundations of organization and institution, and this will require us to have an economics of information and a more sophisticated treatment of the technology of transacting. In the meantime, otherwise good economists will continue to make meaningless assertions to the effect that there is some degree of failure in all real markets because of the ubiquitous presence of transactions cost, or to the effect that in certain situations (characterized by prohibitive transactions cost) the problem is one of "not enough markets." As long as market failure is defined in terms of failure to achieve costless Pareto-efficient allocations, every real world institution, whether decentralized or centralized, is in danger of "failure." Fortunately for the economy, but unfortunately for academic economics, this formulation of the problem of Pareto efficiency is not the problem that real markets and other allocative institutions attempt to solve.

I also believe that, as an intellectual movement in economic theory, radical

The State of Theory

By HARRY G. JOHNSON*

There is, I take it, widespread agreement among both leading economic theorists and economists generally that the role of theory in relation to the subject is at present both ill-defined and of controversial relevance and usefulness. It is worth remarking to start with that this situation can be expected to an important degree to be endemic. In the eighteenth and nineteenth centuries economics *was* theory, in the sense that its main endeavor was to explain precisely the systematic nature of the organization of the economy, a task made difficult by the substitution of the apparently unplanned market for the mercantilistically conceived nation-state. And economics remained theory to an important extent in the first third, and especially the fourth decade, of the twentieth century, though with the major exception of Keynes's *General Theory* theoretical advance consisted already of consolidation rather than revolution—as evidenced by the development of imperfect competition (recognition of the changing industrial framework), the new welfare economics (elimination of a basic logical error in the assumed scientific basis of ethical judgments) and the theory of general equilibrium (recognition that the trained theoretical mind did not need to water down general interdependence into the businessman's pabulum of *ceteris paribus* assumptions).

But consolidation has meant both that an increasing proportion of the profession has been concerned with acquiring expert knowledge about a small piece of a system assumed to be in broad outline a well-known and going concern (and therefore is inevitably driven back into *ceteris paribus* assumptions), and that specialist theorists have had to demonstrate their worth either by the armored might of capture of exceptional cases and the clever trapping of tiny specimens of behavior hitherto considered too unnourishing or insufficiently pestilential to be worth worrying about, or else by the use of ever more elegant mathematics to arrive at ever more economical statements of general equilibrium results.

This situation, as mentioned, must to some extent be endemic, given the high traditional prestige of theory in the subject, the diminishing returns to specialized theory once the main theoretical corpus has been blocked out, and the demands for application of a demonstrably useful science to pressing problems. It is exacerbated by doubts on both sides about the usefulness of theory as largely practiced at present, by the fact that the value of a theoretical contribution is closely inversely correlated with its accessibility to the applied economist, and by the fact that it still seems to be true that professional status can be more quickly and lastingly established by a theoretical article than by a piece of applied research—since the journal editor and his referees, rather than the readers of the journal, are likely to be the final arbiters of competence for theoretical articles. But it has become particularly acute in recent years with the radical rejection of capitalism—in which mathematics has frequently been identified with apologetics—and with the swing of concentration of economic writing and research from "objective" problems to problems like

---

* University of Chicago and London School of Economics and Political Science.

health care, the environment, and poverty and equality, where problems of common sense, historical perspective, and balanced social philosophy matter relatively far more than both sophisticated theory and conscientious number-grubbing.

In this situation there has arisen, not exactly a crisis, but a dangerous gap between "theory" and "applied economics," a gap into which what used to be called "political economy" has sunk with barely a ripple. I would call this gap dangerous especially because the resources to fill it, while qualitatively good enough, may well be too blinkered in imagination to rise to the occasion, or may even do more harm than good. To indicate very briefly what I mean, from different angles: we have a country in which a large-scale demographic shift has produced a population largely rooted neither in the east coast Europe-oriented regions of long-standing British-tradition-dominated established culture, nor in the pioneering near-western regions of fundamentalist simplicity and simple homely virtues, but instead essentially restless, nomadic, and culturally superficial. We have a profession still crowned by an elite of old New England universities, in whose conception of the world Washington and a vague sort of patrician liberalism are enshrined, but the majority of its members are professionals in economics and suburban householders otherwise. We do have among our ranks many still versed in the literature and history of political economy as one of the social-philosophical subjects necessary to the understanding of man in his society, but almost all of these are college teachers and most derive their roots from Europe rather than America. Above all, we are conditioned by the American myth that society started today and will continue the same tomorrow.

Well, what is to be done about it? Not much, on the whole, but some things might help. Senior professors might be encouraged to give a semester course surveying theory, not in technical terms, but in terms of the main concepts and problems that theorists have coped with over their own professional lifetimes. More technical, but still survey-type, courses could also be provided. These would be aimed at providing a basic understanding of such simple concepts as interest and discounting, income and capital, the budget constraint on choice, the cost and time-dimension of adjustment to change, the difference between social and private costs and benefits, and similar notions and would emphasize the many apparently strange contexts in which the essentially same concepts are the starting point of effective analysis. The basic problems of economics are simple; the hard part is to recognize simplicity when you see it. The next hardest part is to present simplicity as common sense rather than ivory tower insensitivity. Theory needs to teach more of both.

Is the Theory of Competitive Equilibrium With It?

By TJALLING C. KOOPMANS*

Recent discussions in the literature have raised serious challenges to the theory of competitive equilibrium. Looking at the debate, I perceive two main issues. I put them in the form of two questions to which, by way of advance summary of my remarks, my own tentative answers are attached.

One. Does the model of competitive equilibrium (the "*CE* model") in its simplest form represent one useful pure and special case, one valuable foothold for a steep climb? My answer: Yes.

Two. Can we as yet evaluate the merit or promise of the various ramifications of the theory in recent literature? My answer: I find it hard to assess this fascinating blend of high achievements, challenging starts and possible dead-ends.

I shall mix the motivation of my answers in with comments on some recent criticisms. I am thinking in particular of John Kenneth Galbraith, Nicholas Kaldor, Janos Kornai and Martin Shubik. Since Professor Galbraith is with us, I shall not try to anticipate him.

Beginning with question one, I think the issue does not lie in the mathematical form of the theory. The contributions made by mathematical reasoning were necessary if the problems put were to be answered. The issue is the problems put. I do not hesitate about my "yes" to question one, because of the great value to economic theory of a fully worked out special case. I entirely agree with Kornai about the many aspects of reality ignored in the *CE* model which are brought out in his book: the importance of the control system in all existing economies, the role in that system of information about quantities as well as prices, the importance of increasing returns to scale, etc. I differ only on what to do from here on. Kornai's "revolutionary" proposal is now to start afresh with entirely new approaches embodying other aspects of reality, which he enumerates with care and perception. The "reformist" alternative, which Kornai rejects, is to amend and extend the given special case by grafting other important aspects on to it. I think *both* should be attempted.

Kaldor also calls for "a major act of demolition" of the basic conceptual framework of equilibrium theory, without which "it is impossible to make any real progress." The case for the reform approach can be illustrated by what seems to me Kaldor's gravest charge: that equilibrium theory ignores the pervasiveness and inexhaustibility of increasing returns to scale. I think that this important point can be met at least half-way, by introducing further assumptions that bear on the way time and space enter into the problem. As to time, the construction of the next generation of capital goods for larger scales of processing requires time. Kaldor recognizes that the allocation of the existing capital stock and other factors to current production can meanwhile be administered by short-run equilibrium prices. A recent paper by M. L. Weitzman for a two-sector model leads me to expect that an efficient stringing together of a sequence of such temporary equilibria is mathematically feasible in the presence of increasing returns, at least in terms of an "optimal growth" model. If so, this would support an allocation scheme on the "Hungarian plan," in which investment in large capital units is centralized, while capital utilization decisions are decentralized in a price-guided way. However, extension to a complete intertemporal market model might be subject to the same realistic limitation as that noted by Kaldor in regard to exhaustible resources: Firms might not exercise foresight long enough ahead for the time scale required by the problem. A relevant spatial factor is the cost of transporting the product from producer to user. This will slow down the increase in scale, and may even balance it out in an ultimate equilibrium for a number of goods.

---

* Yale University.

Both Kornai and Shubik stress the need for a more detailed modelling of the circumstances and constraints under which suppliers and demanders obtain access to each other and conclude their trades. In particular, Shubik is concerned with constraints on credit that are expressed by a liquidity requirement constraining an agent at every moment of time. This is in contrast to a budget constraint applicable just to the sum of expenditures minus receipts over an accounting period. It will require modelling the economy as a process, a sequence over time of moves fully feasible to the individual acting by himself. In the presence of uncertainty this should include bankruptcy proceedings as a possible outcome for the agent. Shubik may well be right in his claim that this model cannot be accommodated within the framework of the *CE* model. However, logical links are undoubtedly present.

The asperity of some of the criticisms seems to me in part provoked by the spill-over into mathematical economics of attitudes and traditions imparted by training in mathematics. Mathematical economics, while long on rigor, has generally been short on interpretation and practically silent on motivation for its choice of problems. Thus, the sheer volume of work expended on the theory of equilibrium is taken by the general economist as evidence of a strong belief in the explanatory reach of the model. I think the explanation is different. While I was searching for words to express the phenomenon, the precise words were presented to me by David Freedman, mathematician, probabilist and statistician, in a recent seminar at Yale: "Mathematicians are incapable of leaving well enough alone." In other words, any particular problem or model should be plumbed to its full depth and given its most general formulation. As a result, the shifts of focus, the breaking out of molds needed from time to time in any empirical science, take longer to come about.

The counterweight to this element of mathematical style must and has come from economists with a more pragmatic outlook. There is a continuing need for general and critical discussion of the choice of problems to be examined in economic theory. By its nature, rigor has no hold on this question. While the criticism takes shape, substantial further progress has been and is being made within the *CE* model, due also to its great flexibility of interpretation. I shall expand on my answer to question two in a brief discussion of this work. The task is facilitated by the important recent book by Kenneth J. Arrow and F. H. Hahn. It breaks a tradition by being generous with interpretation and by providing in the preface a motivation that is specific, frank and subtle. Important to the general economist is its careful discussion of present knowledge about the uniqueness and stability of equilibrium. The validity of the widely used method of policy analysis by "comparative statics" is found to be as yet subject to qualifications.

Another innovative recent book, by Herbert Scarf (with the collaboration of T. Hansen), enables us to make such applications with greater precision and to cases involving more variables and relationships. The algorithm approximates an equilibrium as a fixed point of a continuous mapping. An application to a problem of taxation has been made by John B. Shoven and John Whalley.

Two earlier developments are in the nature of magnificent *tours de force*, enriching our insight, but with a somewhat strained relation to reality. One of these, originated by Arrow (1953) and extended by Gerard Debreu (ch. 7), introduces uncertainty by considering trade in commodities contingent on the "state of nature." The approach is a heroic attempt to stretch the *CE* model as far as it will go. Roy Radner (1967, 1968) examines the limitations of the attempt connected with the different information requirements placed on the agents.

The other *tour de force* concerns the connection between the set of competitive equilibria and the game-theoretical concept of the core of an economy when the number of agents increases without limit. Roughly, an allocation is in the core if no coalition of any number of agents can do better for its members by coordinated quantitative bargaining. The principal result, suggested or proved with increasing generality by Shubik, Scarf, Debreu and Scarf, and Arrow and Hahn (ch. 8), is that the larger the number of agents, the closer any outcome of core-type bargaining comes to some competitive (price-guided) equilibrium. A fanciful extension by R. J. Aumann (1964, 1966), Karl Vind, Debreu, W. Hildebrand, Arrow and Hahn (see ch. 8) and others utilizes the concept of an economy with an infinity of agents, whose characteristics are continuously distributed in a measure space.

The interpretative weakness of both *tours de force* lies in the information handling requirements implicitly placed on the agents without regard to cost or even feasibility. In the case of contingent trading, these requirements go up in proportion to the number of states of nature, hence much more steeply than the number of future periods beset with uncertainty. With regard to the concept of the core, they go up in proportion to the number of coalitions containing a given member, that is, much more steeply than the number of members. It seems to me that these two lines of work will either remain standing as essentially unused but brilliant interpretational or mathematical feats—or one or both may possibly be the beginning of a long development in which institutional detail is introduced piece by piece, to represent the limitations of markets that deal with the future, or the various barriers on coalition formation existing in society.

I conclude with a few remarks about the finite competitive economy that assumes objective certainty about the outcome of given actions by all agents. There has been an increasing concern in the recent literature with fitting adjustment processes into that model. I attribute this concern to a weakness in the model—not a logical but an interpretative weakness. Optimizing responses of economic agents are simultaneously feasible only if the proper prices are already known to them. But these prices must somehow themselves be the result of these same responses. Thus there is something circular in the description of events. The market participants must be endowed with extrasensory perception (if acting simultaneously) or with supernatural premonition (if acting successively). One is thus led to think in terms of a process in which information flows, the individually feasible actions they induce, and the new flows generated by the actions, etc., are spelled out sequentially. I believe (with Kornai and Shubik) that such processes are now the more challenging objects of research, for the following reasons:

1. The notion of a process does not presuppose the approach to an equilibrium.
2. A process that does not approach an equilibrium (under constant technology and preferences) can be of great interest in itself.
3. An equilibrium that is not approached by any process that starts from a different initial state is of no interest in itself.
4. With the process notion, one can relieve the strain on our credulity and experiment with diminished degrees of individual rationality, perception, foresight, capacity for computation, and formalization of states of uncertainty. We could then at least ask how much of such relaxing of rationality still allows an equilibrium with some optimality properties to be approached.
5. In a world of continuing but only dimly foreseeable change in technology and in preferences, the notion of equilibrium disappears, but that of an adjustment process remains.

Possible connections between processes and equilibrium can be glimpsed from some recent work in "optimal growth" theory. One can look at the study of optimal growth as a scouting device that suggests useful conjectures for more complicated market processes which do approach an optimizing equilibrium. Some pointers of this kind are contained in work on many-sector models by W. R. S. Sutherland and by T. Hansen and Koopmans, and in analogies with known properties of one- and two-sector models. The "invariant capital stock," which generalizes to n dimensions the "golden rule stock (modified by discounting)," has a mathematical structure similar to that of a competitive equilibrium. It is a "fixed point" of a mapping and can be computed by fixed-point methods. This approach may therefore allow a reconnaissance of the problem of distinguishing interesting equilibria from uninteresting ones.

REFERENCES

K. J. Arrow, "Le Rôle des Valeurs Boursières pour la Répartition la Meilleure des Risques," *Econométrie*, 1953, 41–48, repr. in Eng. trans., as "The Role of Securities in the Optimal Allocation of Risk-Bearing," *Rev. Econ. Stud.*, 1963–64, *31*, 91–96.

——— and **F. H. Hahn**, *General Competitive Analysis*, San Francisco and Edinburgh 1971.

R. J. Aumann, "Markets with a Continuum of Traders," *Econometrica*, Jan. 1964, *32*, 39–50.

———, "Existence of Competitive Equilibria in Markets with a Continuum of Traders," *Econometrica*, Jan. 1966, *34*, 1–17.

G. Debreu, *Theory of Value*, New York 1959.

——— and **H. Scarf**, "A Central Limit Theorem on the Core of an Economy," *Int. Econ. Rev.*, 1963, *4*, 235–246.

J. K. Galbraith, *Economics and the Public Purpose*, Boston 1973.

T. Hansen and **T. C. Koopmans**, "On the Definition and Computation of a Capital Stock Invariant Under Optimization," *J. Econ. Theory*, Dec. 1972, *5*, 487–523.

W. Hildebrand, "On Economies with Many Agents," *J. Econ. Theory*, 1970, *2*, 161–188.

N. Kaldor, "The Irrelevance of Equilibrium Economics," *Econ. J.*, Dec. 1972, *82*, 1237–1255.

J. Kornai, *Anti-Equilibrium*, Amsterdam 1971.

R. Radner, "Equilibre des Marchés a Terme et au Comptant en Cas d'Incertitude," *Cahiers d'Econometrie*, Nov. 1967, *12*, 35–52.

———, "Competitive Equilibrium under Uncertainty," *Econometrica*, Jan. 1968, *36*, 31–58.

H. Scarf, with the collaboration of **T. Han-**

sen, *The Computation of Economic Equilibria,* New Haven 1973.

J. B. Shoven and J. Whalley, "A General Equilibrium Calculation of the Effects of Differential Taxation of Income from Capital in the U.S.," *J. Pub. Econ.,* 1972, *1.*

——— and ———, "General Equilibrium with Taxes: A Computational Procedure and an Existence Proof," *Rev. Econ. Stud.,* Oct. 1973, *40,* 475–490.

M. Shubik, "Commodity Money, Oligopoly, Credit and Bankruptcy in a General Equilibrium Model," *West. Econ. J.,* Mar. 1973, *11,* 24–38.

W. R. S. Sutherland, "On Optimal Development in a Multi-Sectoral Economy: The Discounted Case," *Rev. Econ. Stud.,* Oct. 1970, *37,* 585–589.

K. Vind, "Edgeworth Allocations in an Exchange Economy with Many Traders," *Int. Econ. Rev.,* 1964, *5,* 165–177.

On Stabilization Policy: Goals and Models

By THOMAS MUENCH AND NEIL WALLACE*

Ordinarily various fields within economics are distinguished from one another by the phenomena they seek to explain. Macroeconomics and monetary theory are different. While there are special phenomena they seek to explain—fluctuations in aggregate output, employment and the price level for macroeconomics, the structure of markets and the mechanism of exchange for monetary theory—these fields are also distinct from other branches in the methodology they employ. Macroeconomics and monetary theory have traditionally made use of models that contain behavior relationships not grounded explicitly in theories of individual optimizing. And this feature explains why there are associated with these fields kinds of welfare economics based on *ad hoc* notions of desirable time paths of aggregates like unemployment and the price level. Standard welfare economics based on Pareto optimality can be applied only to optimizing models.

Virtually no one, we think, regards this situation as desirable. Macroeconomics developed this way because there did not exist optimizing models that could account for some obvious facts—the most important being fluctuations in aggregate real economic activity. Monetary theory developed this way because there does not exist a theory that describes the way financial institutions and the mechanism of exchange develop and operate under various government rules.

Our theme here will be the contrast between the implications for stabilization policy of traditional *ad hoc* macroeconomic and financial sector models and the implications of a newly developing microeconomic approach based explicitly on optimizing behavior. For the most part, we shall review existing literature both with respect to a critique of the use of *ad hoc* models and with respect to new microeconomic approaches. We shall, though, attempt to indicate fruitful avenues for further microeconomic study of the phenomena and goals that have traditionally fallen within the domain of macroeconomics and monetary theory.

I. The Traditional Macroeconomic Policy Model

One might describe the essentials of the process by which policy decisions about stabilization goals are currently made accurately, albeit in caricature, as follows: The decision-making body uses a model to project the effects over the next year or two of alternative time paths of the

---

* Staff members of the department of economics, University of Minnesota, and consultants to the Federal Reserve Bank of Minneapolis. We wish to thank our colleagues John Kareken, Thomas Sargent, and Christopher Sims for helpful discussions of some of the points raised in the paper. However, all views expressed are the sole responsibility of the authors and should not be interpreted as representing those of the above named colleagues, the Federal Reserve Bank of Minneapolis, or the Federal Reserve System.

policy instruments, and then chooses one of the alternatives.

The fact that policy is conducted this way is in part explained by two properties of the kinds of models used First, the models are not derived explicitly from a theory in which agents optimize subject to constraints. Second, the models treat some version of the observed negative correlation between unemployment and either the level or rate of change of money wage rates or of prices as a structural relationship. The first property makes it impossible to apply standard welfare economics based on Pareto optimality so that if a formal decision or control theory approach is to be followed, then some *ad hoc* criterion must be adopted. If the criterion adopted involves both unemployment and the price level—as would seem to be dictated by the Employment Act of 1946—the second property makes the decision problem more difficult. When treated as a structural relationship, the Phillips curve implies tradeoffs both over time and during any period between the level of unemployment and the time path of the price level.

Despite this difficulty, the problem can be treated in a control theory context if an acceptable criterion function and an acceptable model of the economy can be found. The models which are generally thought to qualify consist of nonlinear stochastic simultaneous equations in which only a finite number of lags enter, and whose parameters can be considered to be subject to estimation uncertainty. The objective function is often a discounted sum of quadratic losses. Unfortunately, it is not yet possible to find numerically the optimum policy which should be followed under such specifications, for methods of finding the optimum lie just beyond the frontier of known techniques. (See Jacques Drèze, especially the end of section 6, and Edward Prescott.) It is known, however, that this framework for macroeconomic policy leads to an adaptive quantitative rule. By quantitative we mean there is generally an optimum policy which depends on the values of the model's parameters. By adaptive we mean that the actual values for the instrument variables depend on recently observed values of the endogenous and exogenous variables.

But can a defensible criterion and a defensible model from among existing estimated macroeconomic models be supplied? We think not. The control theory approach requires that the model be accepted as a structure that will continue to hold in the face of alternative rules for the policy variables: the model's parameters may be imperfectly known, but their distributions cannot depend in unknown ways on the rule adopted. If existing estimated models satisfied this requirement, such models would survive standard statistical tests for structural change. But we know only of models that have failed such tests. (See Thomas Muench et al.).

Why do such models fail? Robert Lucas' (1973) convincing argument offers one reason. His view is that existing macroeconomic models consist largely of estimated relationships which are thought to follow from optimizing theories. When closely examined, however, these theories can be seen to imply that the observed relationships or correlations will change in a systematic way when government policy rules change. The models are defective for the analysis of alternative policies because they treat as structure the observed correlations rather than the underlying optimizing theories. More specifically, the kind of models generally thought to be relevant to the study of macroeconomic phenomena are those in which agents optimize over time in a stochastic environment. In such a setting, actions can be thought of as depending on the agents' views about the future, which in turn must depend on the available information. But if these agents are using information optimally, the way their views about the future depend on that information is influenced by their notions of how government actions depend on the information. When government behavior changes, the public will in general alter the way its views about the future depend on the available information variables. The result is a change in observed correlations.

Lucas' view allows us to analyze the effects of government actions only on those aspects of behavior for which there is a well-developed optimizing theory. Until recently, this was not the case for labor market behavior and, in particular, for the observed Phillips curve. Since one's interpretation of the Phillips curve is central to one's views on stabilization policy, it is not surprising that there is widespread disagreement about what stabilization policy can and should accomplish. Happily, we seem now to have some microeconomic models that can explain the Phillips curve.

II. Microeconomic Approach

We view the microeconomic approach to macroeconomic policy as having the following characteristics: (1) the models are based explicitly on the assumption that individuals optimize; (2) they can explain observable phenomena such as the Phillips curve, unemployment (voluntary and involuntary) and inflation; (3) the models allow for the analysis of a range of policies; and (4) the traditional (Pareto type) welfare criterion is applied to evaluate policies.

In the rest of this section we discuss several approaches that have been made or might be made in the above outlined microeconomic direction. In these discussions one question keeps presenting itself: Are there realistic models for which an adaptive quantitative policy is optimal, or can optimal policies be stated in qualitative terms only (such as, increase the money supply at any fixed rate)?

Money, the Phillips curve, and welfare. In a recent article, Lucas (1972) provides an interesting model which can be used to investigate simple policies dealing with the rate of increase of the money stock, aggregate output, the price level and welfare. In this model there is a perpetual succession of generations of individuals who live two periods. They work the first period, when "young," consume a portion of what they produce and save the monetary proceeds from selling the rest for their "old age." The driving mechanisms are a random division of the younger generation into two market segments (so that participants cannot, in general, infer in which segment they are located) and a random distribution or levy of money on the older generation which is proportional to their first-period holdings.

Lucas shows (to an approximation) that, when the money transfer is actually random, there is a statistical Phillips curve, but that policymakers cannot exploit the tradeoff described by the curve in the sense that if the transfer is fixed and, hence, known, the distributions of real output and utility valves are invariant over different values of the transfer. (In fact, then, all real values for an individual depend only on the size of the market segment in which he is located.)

Let us examine this conclusion in more detail. The Lucas model produces a positive correlation between the rate of change of the observed money stock and total output. Why does this correlation not give the government control of total output? It would have control only if it could change the value of the money stock *relative to* that implied by the scheme that the public thinks the government is pursuing. In other words, for the government to have control, it must be able to create deliberately an unforeseen aggregate disturbance in a particular direction. But since the public is aware in a rough sense of the concerns of the stabilization authorities and is also aware in a rough sense of the basis on which they make decisions, that may not be possible. The stabilization authority can take an action more or less secretly, but it cannot be at all confident about the direction in which it is fooling the public. In fact, such behavior on the part of the stabilization authority amounts to the adoption of a random rule like that analyzed in Lucas' model—under which no effective control over the economy can be exercised.

But what are the welfare implications of a nonrandom money transfer, and what other types of policy might be available in this model? For Lucas, a policy must be uniformly better over all contingencies for each individual if it is to be Pareto superior. This is equivalent to applying the standard Pareto criterion "after the fact" so that individuals are treated as different if the size of the market segment in which they are located is different either when young or when old. Adopting this definition, Lucas shows that a fixed money transfer is Pareto optimal with respect to allocations that can be expressed as functions of the market size of a segment. Moreover, it appears to us that a random transfer would be Pareto inferior to some allocation based on segment size. These welfare results imply a *qualitative* rule for money transfers, namely, they should be nonrandom and proportional to segment size.

However (and here we depart from Lucas' welfare criterion), we could just as well view all individuals in this model as alike "before the fact" so to speak. We would then judge policies by what they imply for the expected utility of the typical individual. This would lead us to look for policies that more optimally distribute risk between youth and old age. A particularly simple policy—which might be called a social security scheme—would give (or tax) the older generation a flat $k and tax (or give) each member of the younger generation in a segment an equal amount sufficient to pay (or receive from) the older generation its $k. In an appendix (available on request) we show that a scheme of this sort could lead to a more optimal allocation.

Now what does all this mean for stabilization policy? The absence of complete contemporaneous and intertemporal (intergenerational) markets is necessary for the Lucas model to produce a statistical Phillips curve and nonzero desired money holdings. But it is precisely their absence which is also likely to mean that government interference—an example being the social security scheme outlined above—is likely to be beneficial, since such interference could substitute, perhaps imperfectly, for complete markets between generations. Such interference, though, does not resemble traditional stabilization policy in that it does not rest on exploiting an observed Phillips curve.

Voluntary unemployment. Unemployment is another observed phenomenon with which micromodels should deal. The first type we discuss is voluntary unemployment, unemployment chosen by an individual in order to seek a better rate of pay in other markets. This gives rise to the type of search models described in Edmund Phelps et al. If individuals predict rationally, their search leads to a natural or equilibrium rate of unemployment. A well-specified stochastic model in which such an equilibrium is shown to exist has been presented in a paper by Lucas and Prescott.

This model has an infinite (or very large) number of discrete markets, in each of which the demand curve for labor follows a given stochastic process. A worker must be unemployed for one period in order to search in other markets. Search itself is random, but in an expected value sense it finds the best possibilities. An important aspect, which we will discuss below, is that workers maximize expected discounted lifetime earnings.

Presumably, if a monetary sector and an aggregate disturbance were added to this model, a statistical Phillips curve would emerge as before, but it again would not be usable for policy purposes. Also, if expected discounted earnings are the appropriate welfare criterion, the equilibrium appears to be Pareto optimal.

However, the expected discounted earnings criterion seems implicitly to assume infinite (or very long) lived individuals who can freely borrow on the basis of their expected earnings. But since workers are finite lived and may default on borrowing, individuals cannot freely borrow against future expected earnings and this welfare criterion is not the relevant one. Workers could no longer count on a certain consumption stream based on a stochastic earnings stream.

In this case their consumption stream would also tend to be random, presumably lowest when they are unemployed. Given risk aversion, the optimality of the competitive solution depends on what markets exist. In the absence of private markets that allow the risk of being unemployed to be shared optimally, a government-run unemployment compensation (or tax) system might be helpful.

As this discussion suggests, analysis of such questions requires a more completely specified version of the Lucas-Prescott type model. We should stress, though, that we think the major implications of the Lucas-Prescott model are robust. Among these implications are that less unemployment is not unambiguously to be preferred to more unemployment, and that search unemployment can generate a nonexploitable Phillips curve.

Involuntary unemployment. While fluctuations in employment and aggregate output can be shown to be perfectly consistent with purely voluntary processes of job search and intertemporal leisure-consumption—a major accomplishment, we feel—it must be admitted that most of the concern about unemployment arises from its seemingly involuntary nature. Involuntary unemployment seems to arise whenever individuals cannot sell their services at ruling prices—i.e., when layoffs occur or when qualified potential workers are not able to obtain positions at existing wage rates. In one sense, such situations seem to be obvious departures from competitive equilibrium and, hence, from Pareto optimality. But as Armen Alchian has emphasized, nonmarket clearing in the above sense is a phenomenon of almost every market in varying degree. It takes the form of queues and inventories.

In the models we looked at above, involuntary unemployment cannot occur because every time a new random state occurs, either of demand or of potential work-force, a new market clearing wage is determined instantaneously. But this seems to be a quite special case. If negotiating a (new) wage is costly, the stochastic processes involved may be such that it is not even feasible to renegotiate a new wage subsequent to the occurrence of each event. Moreover, if there is necessarily risk bearing, if there is no opportunity to hedge against uncertainty, and if workers differ (say, because they are in different stages of their finite working lives), then the allocation of risk bearing resulting from spot market clearing conditional on each new event will not obviously be optimal. In this connection, it may be worth noting that layoffs typically occur on a seniority basis. From the point of view of workers entering such employment for the first time, there are elements of an unconditional forward contract in that kind of arrangement.

These remarks, of course, consist only of hints about the kinds of models that might generate involuntary employment. Nevertheless, some tentative conclusions seem to emerge. Since a general model of the operation of markets is unlikely to imply that it is nonoptimal for there ever to be goods waiting to be sold or orders waiting to be filled, it is also unlikely that a model that generates involuntary unemployment will imply that less involuntary unemployment is always preferable to more involuntary unemployment. And, as long as such models assume optimizing behavior and hence rationality, it seems unlikely that they will imply the existence of an exploitable Phillips curve.

Control of the price level and the financial system. Monetary theory has traditionally been concerned with how money ought to be distributed, i.e., with the kind of monetary standard which ought to be adopted and the kind of regulations on financial institutions which ought to be imposed. It has also been concerned with how much money ought to be distributed, or (if money is being distributed in such a way that its growth rate matters) with what its growth rate ought to be. In this section we briefly consider some models that have been proposed for dealing with these questions.

What we would like to have, of course, is a single model of optimizing individuals in which one could analyze these questions and those concerned with unemployment and output. Unfortunately, we do not have models in which we can satisfactorily deal even with the monetary theory questions alone. Within the monetary area, we start with models of the optimum growth rate of the money stock.

The basic assumption of these models is that money is distributed so that accompanying alternative growth rates of the amount distributed and the implied alternative growth rates of the price level are alternative real rates of return on money. In order for this to be the case, it seems to us that individuals must differ; if everybody is the same and is treated the same, then any method of distributing money ends up being proportional to each person's holdings. And, since people can associate the implied price-level path with the amount distributed, the perceived real rate of return on money will be unaffected by the amount distributed. Thus, if the government does in fact have control of the real rate of return on money, then alterations in that rate necessarily alter the distribution of welfare among individuals.

The concern of monetary theorists has not been primarily the income distribution effects of alternative real rates of return on money, but rather the deadweight loss effects. The only explicit models of these effects are inventory models. The most ambitious of these is, perhaps, Edgar Feige and Michael Parkin's attempt to extend the William Baumol-James Tobin inventory model to encompass costly exchange in consumption commodity stocks.

In their model the individual allocates wealth among money, income-producing capital, and a stock of consumption goods out of which he consumes at a constant rate per unit of time; he will hold a diversified portfolio because he must pay an exogenously fixed fee for each asset transaction. In this model, it can be shown, there is a socially optimal yield on money.

While suggestive, the Feige-Parkin analysis suffers from several defects. First, it ignores the fact that the number of transactions must be some round number, which turns out to be important. (See Preston Miller.) Second, it ignores differences among individuals. In their model, the price level is held fixed with the government choosing an interest rate on money, the payments being financed by a tax. But if all individuals are identical, the levying of alternative interest rates and taxes would not be viewed as implying alternative rates of return on money. Finally, and most serious, Feige and Parkin are vague about the institutional setting of their analysis. Moreover, even if their framework were complete, one wonders whether the problem of finding the optimal yield on money can be treated adequately within a model that takes the structure of markets, the kinds of assets that exist, and transaction costs as exogenous.

This brings us to the other concern of monetary theory, which has to do with how money ought to be distributed or with the kinds of monetary institutions that ought to be adopted. The questions here range from the nature of a medium of exchange and whether some nonmarket compact is necessarily involved, to the kind of monetary standard that ought to be adopted, to details of the regulation of institutions within the present institutional setting. These questions have traditionally been a stabilization policy concern because of the presumption that different monetary institutions convert exogenous uncertainty into different distributions of aggregate demand.

The nature of an exchange medium is coming under increasing scrutiny. Joseph Ostroy shows explicitly how exchange is facilitated by the presence of an outside agency that enforces solvency by way of bookkeeping in terms of money. Martin Shubik shows that the presence of an exchange medium makes it possible to describe multicommodity exchange as a noncooperative game. But, on the crucial question of what the government's role ought to be with respect to providing a medium of exchange, we seem not to have progressed beyond Milton Friedman, who suggests that the provision of an efficient exchange medium may be a natural monopoly and that, therefore, government regulation may be called for.

If government regulation is necessary, what form shall it take? The virtues of various kinds of commodity standards have long been debated. The debate has been concerned mainly with the production technology of the proposed commodity and with the costs of maintaining and adding to inventories of it. (See, in particular, James Buchanan and Friedman 1953.) The production technology is thought to determine how the system transmits and converts exogenous uncertainty into aggregate demand uncertainty, the claim being that the more elastic is labor demand in the production of the commodity standard, the better the system works. But since these discussions do not take place within explicit stochastic models, it is hard to evaluate this claim. Moreover, most discussions deal with a commodity standard in which the price level, expressed in terms of non-interest-bearing currency, is fixed. Perhaps the government can choose the real rate of return on the currency by its choice of an announced price-level path for the commodity.

Models developed to answer questions about financial sector regulation within the present institutional setting are equally unsuitable. Most financial sector models consist of asset demand and supply functions which at best purport to describe how changes in mean yields on assets affect desired portfolios. However, the models which presumably underlie these functions are stochastic inventory models or mean-variance models, which give a prominent place to other moments of the distribution of returns. According to them, alterations in second and higher moments of the distribution of returns can be expected to shift those demand curves, but in ways not made explicit in the estimated models. Since the alternative ways of intervening in the financial system are expected to affect second and higher moments of the distribution of returns, it seems clear that we cannot analyze such changes within existing models. More fundamentally, existing models take the set of assets and the structure of markets as given. But, again, if financial regulations do anything, they tend to alter the sets of assets in existence and the structure of markets.

III. Conclusion

From what we have said, economists seem some distance from being able to analyze problems in this area in such a way as to lead to quantitative implications for policy. It is interesting that the basic models that seem to be required are those that delve into the structure of exchange—into what markets exist and into how they function. This is an area we are only beginning to penetrate. (See Roy Radner.) When we do, though, we may find that we get a payoff both in terms of the monetary theory questions we have raised and in terms of the questions concerned with fluctuations in aggregate output and unemployment, both voluntary and involuntary.

In closing, we would like to emphasize three points in our discussion: (1) the Phillips curve tradeoff that was thought to be exploitable is lacking in foundation; (2) *ad hoc* criteria, such as less unemployment is better than more, should be reexamined; and (3) useful stabilization policy might use nontraditional instruments and must be tied to a detailed analysis of the market structure.

REFERENCES

A. Alchian, "Information Costs, Pricing, and Resource Unemployment," in E. Phelps et al., *Microeconomic Foundations of Employment and Inflation Theory,* New York 1970.

W. Baumol, "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quart. J. Econ.,* Nov. 1952, *66,* 545–56.

J. Buchanan, "Predictability: The Criterion of Monetary Constitutions," in L. B. Yeager, ed., *In Search of a Monetary Constitution,* Cambridge, Mass. 1962.

J. Drèze, "Econometrics and Decision Theory," *Econometrica,* Jan. 1972, *40,* 1–17.

E. Feige and M. Parkin, "The Optimal Quantity of Money, Bonds, Commodity Inventories and Capital," *Amer. Econ. Rev.,* June 1971, *61,* 335–49.

M. Friedman, "Commodity Reserve Currency," *Essays in Positive Economics,* Chicago 1953.

———, *A Program for Monetary Stability,* New York 1960.

R. E. Lucas, "Expectations and the Neutrality of Money," *J.E.T.,* April 1972, *4,* 103–124.

———, "Econometric Policy Evaluation: A Critique," *J. of Money, Cred. and Banking,* 1973.

——— and E. C. Prescott, "Equilibrium Search and Unemployment," mimeo., Carnegie-Mellon Univ., Nov. 1972.

P. Miller, "The Transactions Demand for Money in a Three-Asset Economy," manuscript presented at the December 1973 meetings of the Econometric Society.

T. Muench, A. Rolnick, N. Wallace, and W. Weiler, "Tests for Structural Change and Prediction Intervals for the Reduced-Forms of Two Structural Models of the U.S.: The Fed-MIT and Michigan Quarterly Models," Federal Reserve Bank of Minneapolis, working paper 19, 1972.

J. Ostroy, "The Informational Efficiency of Monetary Exchange," *Amer. Econ. Rev.,* Sept. 1973, *63,* 597–610.

E. Phelps et al., *Microeconomic Foundations of Employment and Inflation Theory,* New York 1970.

E. Prescott, "The Multiperiod Control Problem Under Uncertainty," *Econometrica,* Nov. 1972, *40,* 1043–58.

R. Radner, "Competitive Equilibrium Under Uncertainty," *Econometrica,* Jan. 1968, *36,* 31–58.

M. Shubik, "Commodity Money, Oligopoly, Credit and Bankruptcy in a General Equilibrium Model," *West. Econ. J.,* March 1973, *11,* 24–38.

J. Tobin, "The Interest Elasticity of Transactions Demand for Cash," *Rev. Econ. Statist.,* Aug. 1956, *3,* 241–47.

On Optimal Taxation with Costly Administration

By WALTER PERRIN HELLER AND KARL SHELL*

In adopting a set of taxes, governments are influenced by the relative costs of administering and enforcing each kind of tax. Similarly, costs of compliance and transactions for households and firms depend on the mix of taxes used. Yet, in previous studies of optimal taxation (e.g., Peter Diamond and James Mirrlees), administration and transaction costs play no role in the determination of government tax policy. We offer here a first attempt at incorporating these important costs[1] into a formal model of optimal taxation.[2] In what follows, the class of tax instruments to be used is endogenously determined by an explicit model of an economy with an administrative sector, rather than being exogenously given as in previous studies.

We build on the equilibrium-with-transaction-cost literature (Duncan Foley, Frank Hahn 1971) to extend the theory of optimal taxation to account for costly transactions. We introduce a new concept, the administrative feasibility set, which describes the costs of government tax administration. Because of limited space, we content ourselves with describing some basic elements of the model and then employing the model to consider whether maximization of social welfare requires production efficiency, i.e., whether production efficiency is "desirable."[3] When transactions are costly, pure production efficiency is optimal only in very special and unlikely circumstances. We next extend our notion of efficiency to accommodate the tradeoff between pure production efficiency and efficiency in transaction. We then go on to prove a theorem establishing the desirability of efficiency of production-*cum*-transaction plans, but strong conditions are required to achieve this result. Contrary to the conclusions of the previous literature—which omitted the role of government administration costs—our examples and theorems suggest that optimality of production-*cum*-transaction efficiency is very much in doubt for real-world economies.

I. The Model

Households. Household h is endowed with a vector of commodities, $\omega^h > 0$, including labor and human capital services. Households can sell only on the wholesale market and can buy only on the retail market. For household h, the vector of

---

* University of Pennsylvania. Research support from the National Science Foundation and the Fels Center of Government is gratefully acknowledged. We thank Tony Atkinson, Bent Hansen, Mark R. Killingsworth, David Starrett, members of the Stanford University Institute for Mathematical Studies in the Social Sciences summer seminar and members of the University of Pennsylvania seminar on the endogenous theory of policy for helpful suggestions. Proofs of theorems appear in a Mathematical Appendix which is available from the authors on request.

[1] These include costs of enforcing tax laws and costs of complying with tax laws as well as costs of information processing and information transmission.

[2] This is not to say that the role which administrative costs should play in a full theory of optimal taxation has been overlooked; just the opposite is the case. Diamond and Mirrlees, for example, state in their Concluding Remarks: "As economists have been aware, the omitted constraints on communication, calculation, and administration of an economy . . . limit the direct applicability of the implications of this theory to policy problems. . . ."

[3] If production efficiency is desirable, then at the optimum there is no tradeoff between equity and production efficiency.

goods purchased (at retail) is $x^h \geq 0$ and the vector of goods sold (at wholesale) is $y^h \geq 0$. The household consumption vector is then $c^h = \omega^h + x^h - y^h \geq 0$. If households face the vector of buying prices p and the vector of selling prices q, then the budget constraint for household h is: $px^h \leq qy^h$ +(profits distributed to h)−(net direct taxes levied on h). The household is assumed to maximize its strictly quasi-concave utility function, $u^h(c^h)$, subject to its budget constraint.

Production and marketing. Firms purchase wholesale inputs from households and purchase wholesale and retail inputs from other firms. Firms sell retail outputs to households and sell retail and wholesale outputs to other firms. Private firm f ($f = 1, \ldots, F$) maximizes profits at prevailing producer retail prices, $\hat{p}$, and prevailing producer wholesale prices, $\hat{q}$. The firm's production-transaction plan must lie within its given *production-cum-transaction* feasibility set, A^f. Let w^f be firm f's vector of net output on the wholesale market (i.e., output sold wholesale minus inputs purchased on the wholesale market). Similarly, let r^f be the vector of net output on retail markets. Then the production-transaction plan of firm f is feasible if (w^f, r^f) belongs to the set A^f. The firm's profits are thus $\pi^f = \hat{p}r^f + \hat{q}w^f$. For each of the F private firms we assume that A^f is a closed, convex set, so that competition is sustainable. Let A^{F+1} represent the government production-transaction opportunity set. We can think of the government production-transaction sector as "firm" $F+1$, but the government need not maximize profits.

The government. Before taxation policy can be studied, the universe of available tax instruments must be specified. We consider a universe which includes all those instruments featured in the previous optimal taxation literature. Some new instruments are also included.

The government has four basic tax-subsidy instruments at its disposal. In general, each tax used and the rates at which the tax is set will affect the government's administrative cost. The basic instruments considered are (1) commodity taxes on retail markets, (2) commodity taxes on wholesale markets, (3) profits taxes, including firm-specific profits taxation rates and firm-specific licensing fees and licensing subsidies, and (4) lump-sum taxes and subsidies for households. Since labor is one of our commodities, we include proportional income taxes in our list of instruments.

Why do we include lump-sum taxes in this list? After all, it is frequently argued that lump-sum taxes are so costly that they are rarely used by governments despite their advantages. Some kinds of lump-sum taxes (e.g., a tax based on age) may not be costly to administer but the use of the kinds of lump-sum taxes which are likely to be interesting for optimal distribution may require information which is very costly to acquire (e.g., "ability" levels of all individuals). Our point is that whether or not the government will employ lump-sum taxes is a matter which ought to be determined within the model rather than by *a priori* specification.

The government administrative feasibility set. How does one model costs of administering the tax system? Following what we take to be the spirit of the transaction-cost literature, we introduce the concept of the administrative feasibility set, G, which directly and indirectly relates g, the vector of real resources used up in administration, to the government tax instruments employed.

Let $\bar{x}$ be the vector of vectors of household purchases, $\bar{x} = \{x^h\}_1^H$, and similarly, let $\bar{y}$ be the vector of vectors of household sales, $\bar{y} = \{y^h\}_1^H$. If there are M commodities and H households, then $\bar{x}$ and $\bar{y}$ are vectors of dimension MH. Let m be the

H-dimensional vector of lump-sum subsidies to households; n the vector of lump-sum taxes. Let $\beta^{1f} \leq 1$ be the rate of profit taxation for firm f while β^{2f} $(-\infty < \beta^{2f} < +\infty)$ is the licensing fee to firm f. Therefore after-tax profit, α^f, is related to pretax profit, π^f, by $\alpha^f = (1-\beta^{1f})(\pi^f - \beta^{2f}) \geq 0$. The requirement that α^f be nonnegative is made to incorporate limited liability of stockholders into the model and is a constraint on the government in selecting β^{2f}. Let β denote the $2F$-dimensional vector $\{\beta^{1f}, \beta^{2f}\}_1^F$.

An administrative plan of the government is described by the vector $(\bar{x}, \bar{y}, m, n, \beta, g)$. The set G is then the set of all *feasible* government administrative plans. Administrative costs related to lump-sum taxes and subsidies and profits taxes are thus accounted for directly through the entry of m, n, and β in the administrative technology. The costs of commodity taxes, which drive a wedge between consumer and producer prices, are indirectly accounted for through the consumer purchase and sales vectors, $\bar{x}$ and $\bar{y}$. Presumably, the "farther" $\bar{x}$ and $\bar{y}$ are from their *laissez-faire* values, the higher is the cost of administering the underlying commodity taxes (i.e., the higher is g). To relate administrative costs directly to commodity taxes would mean including prices in the vectors belonging to G. We have not yet explored this possibility, nor are we confident that such a formulation would more accurately model the administrative costs of commodity taxation.

In order to fix ideas about our description of administrative costs, we can consider as an example the special case of *laissez-faire*. In terms of our nomenclature, *laissez-faire* is feasible only if the vector $(\bar{x}^0, \bar{y}^0, 0, 0, 0, 0)$ is an element in the set G, where $\bar{x}^0$ and $\bar{y}^0$ are *laissez-faire* equilibrium allocation vectors. That is, *laissez-faire* is feasible only if zero government administration costs are incurred when all tax and subsidy rates are zero. This would not be the case if the administrative costs of government enforcement of private contracts are included in G.

Our modeling of costly administration in terms of the set G is, of course, crucial. The usefulness of the subsequent analysis probably hinges on whether or not the administrative feasibility set is an appropriate way of describing the costs of alternative taxation policies. It is, therefore, worth noting some features of this new ingredient in the general equilibrium model of taxation.

First, the administrative feasibility set is expressed in "reduced-form." The production feasibility set of the classical equilibrium model is independent of tastes and endowments of households. In our model, however, the specification of the set G depends on the economy to which it is applied. Thus, one would expect that the administrative costs of achieving an egalitarian consumer allocation for an economy with small differences in initial household endowment are less than such costs for an economy with great disparities in initial endowments. That suggests that the set G is dependent upon the basic parameters of the economy (e.g., the pretax distribution of wealth). This is an important point since cross-section and time-series estimates of the administrative cost set cannot be made until the model is extended to relate changes in the set G to changes in fundamental economic parameters. Nonetheless, the model as it now stands is useful in assessing questions about optimal tax policy in a given economy as long as the set G is known to the policymaker.

Second, the set G is valid only for a given price convention. For instance, if prices are constrained to sum to unity, then administrative feasibility would be described in terms of an administrative set, G_1. If instead, prices are constrained to sum to two, then another set, G_2, would be applicable. Remember that direct household taxes and subsidies, m and n, and license fees, β^2, are in units-of-account (say, dollars). If (m, n, β^2) were doubled and all prices doubled, *ceteris paribus*, should not costs of administration, g, be unaffected? This avoidance of "*numeraire* illusion" is easy to model, and we should emphasize that it is not a statement about the form of any given administrative set but rather how one set G_1 is related to another set G_2.

Finally, for simplicity, it is assumed that resources devoted to enforcement accomplish perfect compliance, in that all households and firms exactly fulfill their obligations. We feel that our model can be extended to handle imperfect enforcement and compliance; indeed, a substantial public finance literature analyzes what might be called the "optimal degree of enforcement." For example, what level of resources should the government devote to tax audit? The dual strategic question for the taxpayer is: What degree of cheating should be risked? In the present analysis such questions are ignored, for at this stage the extension of the model to cover such considerations would only complicate matters unnecessarily.

Social goals. We assume that the government seeks to set taxation policies and operate government production so as to maximize a strictly increasing, individualistic Bergson welfare function $W(u^1, \ldots, u^h, \ldots, u^H)$. The government is constrained by the behavior of households (maximization of utilities subject to budget constraints), the behavior of firms (maximization of profits subject to production-transaction feasibility), feasibility of government production and transaction, and administrative feasibility of government taxation policy. Of course, it is required that materials balance in the economy, which implies (by Walras' Law) that the government satisfies its budget constraint.

In positing a government which maximizes W, we—and previous authors on optimal taxation—have made a sharp and unrealistic distinction between public agents (who are assumed to seek to maximize social welfare) and private agents (who are assumed to maximize their own utilities). We are aware that in a fundamental sense all agents are private agents. Government bureaucrats are red-blooded people with individual goals in possible conflict with social policy. It is our long-run objective to incorporate such considerations into the analysis, but we will find it convenient at present to follow the custom of making this somewhat artificial distinction between public and private agents.

II. Should Production be Efficient or Inefficient?

Diamond and Mirrlees establish the desirability of production efficiency in a model with costless transaction, costless administration of commodity taxes (the only "feasible" instruments) and constant returns-to-scale technologies. Production efficiency is not necessarily optimal when decreasing returns and consequently positive profits are allowed. Partha Dasgupta and Joseph Stiglitz allow for decreasing returns and show that when firm-specific profits taxes are feasible and the private production sector is already operating efficiently, then it is optimal for the government production sector to operate so as to attain aggregate production efficiency. If there are very large administrative set-up costs in moving to a system of firm-specific taxes, then these instruments will not be selected, and one would expect that production-*cum*-transaction inefficiency may be optimal. In what follows, we generalize and formalize this intuition: Absence of administrative set-up costs is crucial to our efficiency theorems. We argue that these set-up costs (or "spikes") are fundamental to tax administration and are likely to be pervasive in practice. The assumptions we need for our first efficiency theorem follow.

Assumptions. (1) A^f is closed, convex and permits free disposal for $f=1, \ldots, F, F+1$. Convexity of the F private sets allows for decentralization with a price system. Convexity of the government set, A^{F+1}, is assumed because it is known in simpler models (Hahn 1973) that nonconvexity of the public production set can by itself lead to failure of the efficiency theorem.

(2) The production-transaction possibility frontiers (the upper boundaries of the sets $\{A^f\}_1^{F+1}$) are assumed to be differentiable and the possibility frontiers are assumed to exhibit maximality. By this, we mean that every point on the possibility frontier is efficient, i.e., there are no horizontal or vertical flats.

(3) The administrative feasibility set, G, is closed, exhibits maximality of its possibility frontier and exhibits free disposal in the limited sense that if $(\bar{x}, \bar{y}, m, n, \beta, g) \in G$ then $(\bar{x}, \bar{y}, m, n, \beta, g^\dagger) \in G$ for any $g^\dagger \geq g$.

(4) The economy exhibits *weak sensitivity* of gross profits to profits taxes. We must rule out seemingly unusual cases where the indirect effects of profits taxation swamp the direct effects. Weak sensitivity obtains when marginal costs of administering license fees and subsidies are zero or small or when the administrative sector is small by comparison with the total level of economic activity.

(5) A constrained welfare maximum exists.

(6) At an optimum, wholesale and retail consumer prices differ for at least one commodity. This (seemingly weaker) hypothesis replaces the Diamond-Mirrlees assumption of a produced consumer good or nonproduced production input. Alternatively, one could retain the Diamond-Mirrlees assumption.

Efficiency Theorem1. If assumptions (1)–(6) are satisfied, then all optima lie on the possibility frontier of the aggregate production-*cum*-transaction set $A=\sum_1^{F+1} A^f$, i.e., efficiency of production-*cum*-transaction plans is optimal. Furthermore, if the sets A^f can be decomposed into separate transaction and production sets, then pure production efficiency is optimal.

Efficiency Theorem 1 requires further interpretation. First, why is it that we speak of *production-cum-transaction* efficiency rather than pure production efficiency? This is because the model allows for tradeoffs between pure production efficiency and pure transaction efficiency. This tradeoff arises at the firm level and consequently also appears at the aggregate level. For example, if production activities are not separable from transaction activities, firm 1 may be slightly less efficient in production than firm 2, while firm 1 is substantially more efficient in transacting, so that 1 rather than 2 should be in operation at the welfare optimum.

Suppose that assumptions (1)–(6) hold and the private production sector is already operating efficiently. Then, the government production-transaction manager should operate to ensure overall production-transaction efficiency. The reason for this is that costs g are continuous in the allocation vector $\bar{x}$ under our assumptions. That is, small changes in $\bar{x}$ can be accomplished by small changes in g. Therefore, the resources needed to facilitate a (sufficiently small) Pareto-improvement in the consumer allocation vector are always available if the economy is operating in the interior of the set A.

Further suppose that firm-specific prices are also possible but that it is always more costly to administer a system of firm-specific prices than any system in which producer prices are uniform. Then, we have established the desirability of efficiency *within* the private sector.[4] This assumption, however, is not realistic, and consequently our result on efficiency within the private sector is of limited applicability as can be seen from the case of intermediate-goods taxation.

---

[4] See the Appendix available from the authors.

In our model, intermediate-goods taxation leads to producer prices which differ across firms. With a positive tax on a particular commodity, the firms producing that commodity face a lower price than the firms buying the commodity for use as an input. It may well be administratively less costly to tax all transactions in the same commodity at the same rate regardless of destination (whether to a household or to a firm) than to administer a system in which transactions between firms are untaxed while sales from firms to households are taxed. Therefore, intermediate-goods taxes, being less costly to administer, may well be employed at an optimum, so that inefficiency is optimal. This contrasts with Diamond and Mirrlees, who find that intermediate-goods taxation is undesirable because administration is costless in their model.[5]

It is important to emphasize the crucial role of assumption 3 in establishing Efficiency Theorem 1. The restrictions on the geometry of G are seemingly mild—allowing for decreasing and constant returns and most but not all forms of increasing returns. In particular, maximality of frontiers rules out "spikes" or set-up costs. In Figure 1, we exhibit a set G not satisfying assumption 3. The horizontal axis "represents" the vector $(\bar{x}, \bar{y}, m, n, \beta)$, while g is measured on the vertical axis. G is shaded and the possibility frontier of G is shown by heavy ink. The "spike" of Figure 1 might be thought of as representing the substantial reduction of costs due to administering a simple tax program. Consider, for example, a simple across-the-board 5 percent-sales-tax regime which makes some target allocations relatively cheap to attain because of ease of enforcement and computation. The 5 percent system is substantially easier to administer than a system of sales taxes with rates varying over commodities but lying between, say, 4.99 percent and 5.01 percent. Moving to nearby allocations requires a discrete jump in real administrative resources g. These "spikes" or set-up costs are contrary to the crucial assumption of the maximality of the possibility frontier of G and thus provide possible examples of the desirability of production inefficiency.

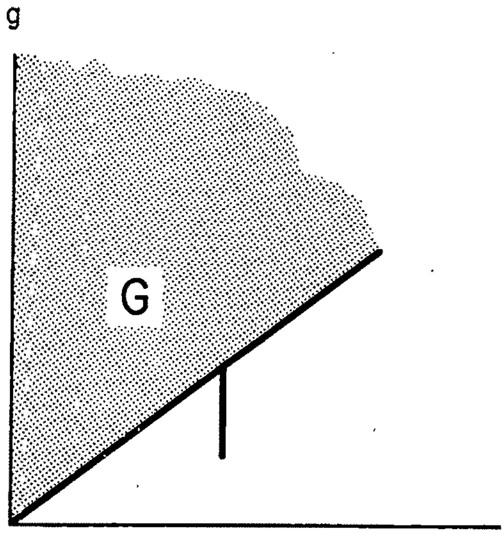


FIGURE 1

Mirrlees puts forward another example in which inefficiency is optimal. An inefficient, decreasing-returns firm is owned by deserving households and thus should be operated because of the profits it distributes to those households. Our efficiency theorem denies the desirability of this inefficiency because of the inclusion of licensing subsidies. The Mirrlees inefficiency example, nonetheless, reappears in our model when assumption 3 is relaxed to allow for "spikes" in G. In particular, if there is a discontinuous jump in administrative costs when the government moves from a system without license fees or subsidies to one with firm-specific license fees and subsidies, then inefficiency may be desirable.

---

[5] Other examples from the policy literature can be cited. Production inefficiency may be desirable in poor countries because tariffs are much easier to administer than income taxes. Similarly, socialism is sometimes thought to be desirable because income redistribution is much more costly under capitalism, even though private ownership may be productively more efficient.

If assumptions (1)–(6) do not hold, it may be the case that the optimum production-*cum*-transaction plans are not efficient. However, at the optimum, costs versus gains are balanced for each of the tax instruments, so we might expect at first blush that the optimal plan would at least lie on the efficiency frontier of the aggregate production-*cum*-transaction-*cum*-administration set (the set of feasible aggregate production-transaction plans net of administrative costs). This is not necessarily the case as the following simple result makes clear.

Proposition. Consider the basic model above except that assumptions (1)–(6) need not necessarily hold. For the economy to exhibit production-*cum*-transaction-*cum*-administration efficiency, it is necessary that both production-*cum*-transaction efficiency and administrative efficiency obtain.

The proposition follows directly from the separability of production-*cum*-transaction plans from administrative plans. So far, there is no tradeoff between production-transaction efficiency and administrative efficiency, although there may be a tradeoff between production-transaction-administrative efficiency and distortions in demand caused by commodity taxation. We can go further and alter the underlying assumptions to consider cases in which production-*cum*-transaction efficiency is not a property of the optimum, but in which production-*cum*-transaction-*cum*-administrative efficiency is a property of the optimum.

One important case is where administrative costs depend directly on production-transaction plans. In discussing Efficiency Theorem 1, we argued that if the economy is operating in the interior of the set A, a feasible Pareto-improving change can be made in consumer allocations because of the continuity of administrative costs with respect to consumer allocations. If, however, it is administratively costly to instruct the government production-transaction sector to seek overall production-transaction efficiency, then the desirability of production-transaction efficiency would be in doubt. This is because the administrative costs in moving to the frontier of A could outweigh the production gains in the move.

Consider another example in which production-transaction plans directly affect administrative costs. Say that firm I is slightly more efficient in production and transaction than firm II, but assume that it is very much more costly for the government to tax firm I than firm II. In this case, it may be desirable to shut down I while encouraging II to operate. In these examples, we are faced with a tradeoff between resources made available by a particular production-transaction plan and the administrative resources used up in accomplishing the particular plan. The next theorem shows that, at the social optimum, this tradeoff is taken into account.

Efficiency Theorem 2. Let assumptions (1)–(3) and (5)–(6) hold except that government administrative costs may also depend on the production-transaction plans of individual firms. Replace assumption (4) with the simpler assumption that marginal costs of administering licensing fees and subsidies are zero or negligible. Under these conditions, even though production-*cum*-transaction efficiency may not be a property of the optimum, efficiency of production-*cum*-transaction-*cum*-administrative plans is optimal.

III. Concluding Remarks

The above analysis is couched in terms

of a "reduced form" administrative feasibility set, G, where the real resources required for administration are directly related to the target levels of the overall allocation vector and thus administrative costs are only indirectly related to the set of commodity taxation instruments. Such direct effects should be incorporated; we plan to do so. We also plan to delve more deeply into the mechanism of taxation. Our thoughts on how to accomplish this are motivated by the following observation. Under a 5 percent sales tax regime, the government acts to encourage the opening of markets in which the 5 percent tax is collected and acts to discourage markets in which the tax is not collected or collected at a rate other than 5 percent. Thus, administrative costs can be related to the costs of government actions in closing down some markets and opening others. Further substantial progress in this area may have to await the development of a general equilibrium theory of markets in which the costs of exchange depend on the mode of exchange.

REFERENCES

P. Dasgupta and J. E. Stiglitz, "On Optimal Taxation and Public Production," *Rev. Econ. Studies*, Jan. 1972, *39*, 87–103.

P. A. Diamond and J. A. Mirrlees, "Optimal Taxation and Public Production," *Amer. Econ. Rev.*, March and June 1971, *61*, 8–27 and 261–278.

D. K. Foley, "Economic Equilibrium with Costly Marketing," *J. of Econ. Theory*, Sept. 1970, *2*, 276–291.

F. H. Hahn, "Equilibrium with Transactions Costs," *Econometrica*, May 1971, *39*, 417–440.

———, "On Optimum Taxation," *J. of Econ. Theory*, Feb. 1973, *6*, 96–106.

J. A. Mirrlees, "On Producer Taxation," *Rev. Econ. Stud.*, Jan. 1972, *39*, 105–111.

ALLOCATING RESOURCES FOR POLICY RESEARCH

How Can Experiments Be More Useful?

By ALICE M. RIVLIN*

Less than a decade ago most economists—save a few mavericks like Guy Orcutt—viewed experimentation as a tool available to physical scientists and psychologists, but not to them. But interest in social experimentation grew rapidly in the late 1960's due at least partly to two developments. First, efforts to rationalize decision making at the Federal level brought economists and other analysts into important government posts and made them painfully aware of how little their traditional data sources allowed them to say about the likely effects of policy changes. Second, the tight Federal budget, the political split between the President and the Congress, and widespread disillusionment with Great Society promises of rapid social change made expensive, new, untested social programs look unattractive even to dedicated liberals and made experimentation seem a more sensible alternative.

Thus far only a few major social experiments have been finished; one might well argue that it is therefore too soon to assess the usefulness of such experiments. But can we afford to wait? Compared with traditional ways of collecting data, experiments are costly and time consuming; and there is clearly some danger that the government will commit substantial resources to experiments that are poorly chosen, designed, or executed and do not yield useful results. The confidence of politicians and the public in a potentially powerful tool could be undermined if experimentation is clumsily used.

In what follows, therefore, I attempt to draw some tentative lessons from the social experimentation experience so far and suggest what kinds of future experiments might be most useful to policy makers. I start from the assumption that resources for experimentation are and will be severely limited, so that it will be particularly important to use experimental resources cost-effectively.

I. Four Types of Experiments

Various social experiments are addressed to distinct kinds of policy problems and raise distinct kinds of design and execution problems.[1] Recent ones appear to fall into four groups, addressed to four types of policy questions.

Type 1

What is the response of individuals, households, or other micro units to a change in economic incentives?

This type of experiment attempts to estimate the effect of a change in a price, tax, or subsidy on the behavior of a large number of independent units such as individuals or households (or conceivably firms in a market with large numbers of small units). The "treatment" is a schedule of prices or taxes or subsidies. The outputs of interest are measures of economic activity (e.g., earnings, hours worked, or expenditures). The units are assumed not to influence each other.

---

* The Brookings Institution. I am indebted to Henry Aaron, James Coleman, Edward Gramlich, Robert Hartman, Guy Orcutt, Charles Schultze, and Michael Timpane for helpful comments on an earlier draft.

[1] Such experiments do have one thing in common: they all raise serious ethical and legal issues analogous to those encountered in biomedical and psychological experimentation on human subjects (Rivlin and Timpane, eds., forthcoming-a).

Examples are: (1) The New Jersey and rural income-maintenance experiments (see Instit. for Res. on Poverty). Concern centered on finding out how hours worked and earnings were related to guarantee levels and the marginal tax rate. (2) The housing demand experiment (see Abt Assoc.). The need was to find out how household expenditures for housing were related to various forms and levels of housing allowance. (3) Health insurance experiments (see Joseph Newhouse). Concern centered on finding out how individual use of medical care relates to the coinsurance and deductible features of health insurance policies. (4) The income-maintenance experiments in Gary, Denver, and Seattle. These are not pure examples of this type of experiment because the participants are offered various types of services as well as money payments.

Problems of execution. Carrying out any social experiment successfully is a managerial tour de force, but Type 1 experiments present fewer problems of execution than other types. The treatment (negative tax schedule, housing allowance, etc.) can be exactly specified. Outputs (earnings, hours worked, etc.) can be identified and reasonably accurately measured. A probability sample of individuals or other micro units can be drawn and units can be assigned to treatment and control groups on a random basis without serious difficulty.

Problems of inference. The validity of this type of experiment depends on the assumption that the behavior of the micro units in response to the treatment is independent. This assumption makes it possible to use a thin sample from a wide area rather than giving the treatment to everyone in a smaller area who would be eligible if the policy were in effect. But, of course, the assumption may be wrong. People may behave one way if they are the only ones on their block receiving a negative income tax or a housing allowance or a government health insurance plan and quite differently if all their friends and neighbors are subject to the same policy. The New Jersey experiment, for all its careful design and execution, may, in fact, tell us nothing about what would happen if all low-income families were eligible for negative income tax benefits.

Even with a thin sample there is still the question of whether to sample from the whole population or from a more limited geographical area. Limiting the area reduces some kinds of variability and greatly reduces the administrative costs of the experiment, but it also increases the chances that exogenous local events will overwhelm the results. This danger was clear and present when New Jersey changed its welfare law in the middle of the experiment and when Seattle's unemployment rate rose precipitously as the local experiment was getting under way.

Two other inference problems that plague Type 1 experiments are common to all experiments with human beings. First, do people behave in the face of a temporary policy in the same way they would in the face of a permanent one? This problem seems most serious in experiments that involve major long-run decisions—like the decision to buy a house or even to move to a different apartment. Some clues as to the importance of the problem may be obtained by splitting the sample in a given experiment, giving one group the policy treatment for a longer time than the other and seeing if their behavior differs, but some uncertainties about the validity of inference from a temporary experiment will still remain.

The second problem is the so-called Hawthorne effect. Do people really behave normally when they know their behavior is being observed?

Type 2

What is the market response to a change in economic incentives?

These experiments relate to the same kind of policy proposals that give rise to Type 1 experiments, but ask a different set of questions. The problem is to find out not how individuals respond to a policy change, but how a whole market responds. How do supply and price, say, of housing respond to a policy-induced change in demand? Or how do demand and price, say, for low-skill workers respond to a change in supply?

Examples are: (1) The housing supply experiment (see I. Lowry, ed.). The purpose of the experiment is to estimate how much housing supply would increase and how much rents and housing prices would rise in response to housing demand generated by a housing allowance. (2) Several similar experiments which have been discussed but not carried out—for example, experiments involving the impact of increased health insurance on the supply of health services; the impact of negative income taxes on wage rates; and the impact of day-care vouchers on the supply of care.

Problems of execution. These experiments are far more difficult to carry out than Type 1, mainly because here it is impossible to use a thin sample. It is necessary to saturate an area, giving the treatment to everyone who would be eligible under a national policy and seeing how the market responds. Saturation experiments are costly, unless the communities being saturated are so small as to be of little national interest. Moreover, they raise political difficulties. Community consent must be obtained, and this may be difficult if the relevant market encompasses several local jurisdictions. Some jurisdictions may refuse to participate; others may try to alter the experiment to their own ends. It may also be necessary to account for the behavior of institutions with market power (e.g., banks) as well as to measure the aggregate behavior of micro units in the market. The outputs themselves may be hard to measure (e.g., housing supply).

Problems of inference. The main problem with Type 2 experiments is that the cost of saturation limits the number of sites and it is risky to generalize about the behavior of markets from samples of one or two. The chances that exogenous local events will invalidate the experiment are even higher than for Type 1 experiments. It might be possible to compare the treatment sites with control sites, but one can hardly put confidence limits on the differences. It is not at all clear how control sites should be chosen or when two cities are really comparable. How would one know whether another city is "like Pittsburgh" in all the dimensions that might affect the results of a community-wide experiment?

Type 2 experiments also raise the familiar uncertainties as to whether responses are affected by the temporary nature of the treatment or by the fact that behavior is being observed. These problems are harder to deal with than in Type 1 experiments, and cannot be remedied, as in a Type 1 experiment, by sample-splitting.

Type 3

What is the production function of a public service, such as health, education, or manpower training?

These experiments arise from uncertainties about a different set of government programs—programs at all levels of government that provide services to people. The hope is that experiments can be used to uncover relationships between inputs and outputs so that the desired outputs can be produced more effectively.

Examples are: (1) Follow Through and Head Start planned variations (see Rivlin and Timpane, eds., forthcoming-b). These attempted to try out a variety of "models" of early childhood curricula and compare the effects of the models with each other and with the effects of regular schooling. (2) Similar experiments under discussion or in some cases carried out on a limited scale concerning health services, child care, family planning, manpower training, and bilingual education.

Problems of execution. Type 3 experiments are even harder to carry out than Type 2. Unlike Types 1 and 2, where the policy change of interest is a relatively simple tax or price schedule, the "treatments" to be compared in Type 3 are complex and hard to describe clearly. They are also likely to result in multiple outputs, and disagreements are bound to arise about how the outputs should be measured and which are most important.

One way of simplifying the problem is to agree on a set of measurable outputs (e.g., reading and math scores of children), then vary easily measurable inputs (e.g., class size) and try to estimate relationships between the inputs and the outputs that can be used to make the system more effective. This approach has been tried frequently in education, but has not yielded many significant results. Class size does not appear to predict test scores of children, which is hardly surprising. Presumably, what does matter is what happens in a class of a given size—what the curriculum is, how the teacher uses time, how he communicates with students, and what he expects of them.

It may be that education experimenters have so far failed to vary the inputs that really matter or to measure outputs appropriately. More mini-experiments should be tried in hopes of finding relationships between inputs that can be manipulated easily in existing schools and desired outputs. On the other hand, it may prove necessary to abandon simple experiments in education in favor of attempts to define more complex programs or models and test these.

The "planned variation" programs were undertaken to compare the effectiveness of disparate and complex models of early childhood education—an objective that proved extremely difficult to carry out in practice. Problems arose because the "models" were not well defined, could not readily be translated into specific directions for teachers to follow, and were changed and developed as the experiment went along. Models were implemented in varying degrees in different sites but no good measures were developed of the extent of implementation. There was little agreement on what output measures were appropriate or how to weight them. For want of better measures, standardized tests were used as outputs, despite the fact that standardized tests are poorly designed for comparing the effectiveness of different curricula and despite strong dissents from the identification of "success" in early childhood education with the acquisition of cognitive skills. Little theory (comparable to the microeconomic theory underlying Type 1 and 2 experiments) was available to indicate the direction, timing, or magnitude of the expected effects of the model curriculum. In addition, some rather rudimentary and avoidable errors made the planned variation results hard to interpret. Comparison groups, for example, were often drawn from less deprived populations than the treatment groups.

Correcting these problems while applying the planned variation approach to other areas would involve (1) developing and pretesting models more fully prior to large-scale experimentation; (2) developing measures of the extent to which a model is actually implemented in an experimental site; (3) developing more discriminating output measures and specifying the direction in which they are expected to move and criteria of "success"; and (4) selecting treatment and control samples from the same population, but not allowing the treatment to spill over into the control samples. All of this requires large resources and a period of several years' time both to develop the models and the measures and to conduct an experiment. It is not at all certain that such an ambitious plan could be carried out in the real world of bureaucracies at several levels of government. And when the answers were finally forthcoming, would they be to questions still considered relevant?

Problems of inference. Two questions about the planned variation approach concern whether one can hold a model of service delivery (e.g., a school curriculum with full specifications about how it is to be implemented) static long enough to test and evaluate it and whether it is even desirable to do so. It can certainly be argued that the *U.S.* system of highly decentralized decision making about services renders invalid attempts to test and compare a rigid set of models. In the real world the models would be altered by those who carried them out at the local level and would end up different from each other and from the original conception. To impose an artificial rigidity for the sake of experimentation can be regarded as stifling creativity and ingenuity.

Another problem is whether the communities (or schools, hospitals, day-care centers, etc.) that participate in an experiment should do so as volunteers. It seems necessary to use volunteers since communities dragooned into participating in a service delivery experiment will probably not cooperate fully or give the models a fair test. But the use of volunteers creates a problem of inference: the results can be generalized only to a population of other volunteer communities, or, indeed, only to communities that would volunteer to implement a particular model.

Type 4

How can incentives be altered in order to affect the production of a service and in turn to affect ultimate outcomes?

These experiments arise out of multistage hypotheses which assert that the way to affect the behavior of large systems (school systems, health delivery systems) is to change the incentives to which the systems respond. Type 4 is a mixture of the first three types and by far the most complicated.

Examples are: (1) Education vouchers (see Rand Corporation). The complex hypothesis behind education vouchers is that giving parents a choice of schools will induce schools to compete for students, which in turn will make schools choose more effective methods and teachers and will alter student outcomes in the directions desired by parents. The current demonstration of vouchers in Alum Rock, California, is limited to public schools, but other tests including private schools are under serious discussion. (2) Health maintenance organizations. Again the hypothesis is complex: that paying health providers for all the care a patient needs in a year (rather than service by service) will induce them to provide a different mix of care, which in turn will improve the patient's health (or lower the cost of the same health result). Tests involve the use of prepaid health-maintenance organizations to provide health care for low-income patients. (3) Performance contracting (see Gramlich and P. Koshel). This experiment grew out of another hypothesis about incentives: that school systems could contract with private firms to teach certain skills, reimbursing the firms on the basis of student performance, and that this in turn will induce the firms to choose effective methods and will result in higher student performance. The actual experiment did not last long enough to test the hypothesis; it showed only that performance contracting did not produce dramatic results in one year.

Problems of execution and inference. Type 4 experiments are the hardest to plan, execute, and interpret. One way to ease the difficulty would be to take a "black box" approach. For example, vouchers could be tried in a sample of communities, and student outcomes in those communities could be compared with outcomes in control communities without any attempt to monitor what happened inside the box—how the changes, if any, took place. But this approach is appropriate only if the main purpose of the experiment is to knock down a hypothesis that changing an incentive structure can have a significant and rapid effect on outcomes. If the government is serious about testing a change in incentives that is expected to be successful and might be widely adopted, it would seem desirable to look inside the black box and try to understand how the change works, what the essential ingredients of success are and how the results are related to changes in the nature of the incentive.

It should be recognized, however, that it may never be possible to try experiments (with multiple sites, controls, and other features required for statistical inference) with such complicated hypotheses. It may be necessary to settle for what can be learned from monitoring the experiences of a series of sites that volunteer to try the policy and adapt it to their own peculiar needs, values, and conditions.

II. Choosing Useful Experiments

One of the big problems of social experimentation is the constant conflict between the pure research objectives of the experiment and other political and policy objectives. One manifestation of the conflict is the desire of the researcher to keep the treatment constant over the life of the experiment, so that he can make estimates of the parameters of some underlying model of the social system involved, and the desire of program operators and politicians to keep changing the treatment, to profit from experience, to increase the chances of a successful outcome, or to avoid public criticism and other difficulties that arise as the experiment progresses.

Researchers conducting Type 1 experiments seem able to stand their ground. To discover individual reactions to changes in incentives, it is clearly necessary to leave the policy constant for an appreciable period. It would not have made much sense to alter the guarantees or marginal tax rates in the middle of the New Jersey experiment; indeed, one can argue that three years of constant treatment were not enough. The goal of finding out how the behavior of individuals is related to policy parameters (e.g., marginal tax rates) can be attained by having additional treatment groups exposed to different values of the parameters, rather than by varying the parameters for the same treatment group over the life of the experiment.

Type 2 experiments, however, are so expensive that it is normally impossible to try multiple varieties of the policy instrument and observe how market responses differ. It is necessary to choose one or two varieties. Then the question is whether to choose policies likely to cause a large shock to the market (e.g., a high-level housing allowance likely to cause a big increase in demand for housing and strong upward pressure on rents) or one likely to cause less disruption. A third alternative is to phase the policy in slowly, endeavoring to minimize the disruption as one goes along. The answer really depends on the political purpose of the experiment—on whether the objective is to find what really happens in the case of a big policy change or to demonstrate that a policy can work smoothly if it is phased in carefully.

Type 3 experiments clearly need a pre-experimental period in which service delivery models are developed and adapted and during which learning from experience is allowed. In the interests of clean research findings, large-scale field tests should probably be delayed until the service delivery models are reasonably stable and able to be replicated, but, of course, it is hard to know when that stage has been reached. Indeed, interest in the innovation may have died before it reaches a stable state. The same observations apply to Type 4 experiments.

Alternatives to Experimentation

Since experiments are expensive and time consuming, alternative ways of getting the information needed for policy purposes—e.g., surveys or data generated by natural-occurring (nonexperimental) events—should be considered in the hope that the needed information can be obtained faster or at less cost.

Rather than conduct Type 1 experiments, one might try to find natural changes in tax, price, or subsidy schedules for some groups in the population (e.g., changes in the terms of employee health insurance in a company) and to observe the behavior of individuals in the face of these changes. Sometimes, of course, no closely related policies exist.

As an alternative to Type 2 experiments, the possibility of finding naturally occurring shocks in comparable markets should be explored. For example, it might be possible to estimate the impact of an increase in demand for housing by observing what has happened in communities that have experienced an upsurge in demand. Such an upsurge might occur naturally where there is a sudden increase in income or employment in a local area.

As substitutes for Type 3 experiments, naturally occurring events are likely to be somewhat suspect—to involve unrepresentative samples or situations especially conducive to innovation in service delivery. Indeed, reliance on nonexperimental methods has produced so few clues to the nature of production functions in the social services that a strong prima facie case can be made for more systematic development and testing of models.

There is not likely to be any alternative to the Type 4 experiment. But it should be recognized that a true experiment is not likely to be possible, either.

III. Prescriptions

Type 1. These experiments have been shown to be feasible and a good case can be made for increased use of them in the future to help policy makers predict the responses of households and other micro units to changes in taxes, prices, and subsidies. Such experiments might be used to help formulate policies with respect to user charges for public services, effluent charges to reduce pollution, higher education tuition and student aid policy, parking fees and other devices for reducing automobile use, student loan policy, as well as income maintenance, health-care financing, housing allowances, and even various forms of tax policy.

However, experimental data would be much more useful if they could be plugged in to the context of more extensive, continuing panel surveys. If the permanent panels advocated by Thomas Juster were available, subsamples of the panels could occasionally be used as experimental groups to test responses to particular policy changes. For example, panels of young people could be followed through the transition from school to work. These panels could also be used to measure responses to various forms of student aid. Subsamples of the panel could be offered student aid on different terms and the differences in their behavior could be measured.

Type 2. These experiments are probably also feasible, but should be used with considerably greater caution than Type 1. Saturation experiments are very expensive and difficult to interpret because controls are not really possible and special local circumstances may confound the results. Hence, they are more likely to be useful as a dry run for a policy whose adoption is imminent than as a way of estimating structural parameters of a market.

Type 3. It has to be recognized that we are not *now* able to carry out experiments with complex service-delivery models. The models are not well enough defined; outcome measures are not well enough developed. The planned variation experiments in education yielded few interpretable results because they were conducted without sufficient effort to define the models and develop measures of outcome and implementation. These mistakes should not be repeated. An experimental strategy may be worth the effort (especially since there are no obvious alternatives), but such a strategy is unlikely to pay off unless considerable time and resources are devoted to development before large-scale experiments are conducted.

Type 4. Field trials of policies to alter incentives of the large institutions (e.g., vouchers in a school system) can certainly be valuable, but it has to be recognized that the innovation will be developing as the trials go along. It may never be possible to do "experiments" on which firm statistical inferences can be based.

IV. Conclusions

The cautionary tone of this paper should emphatically not be construed as a rejection of experimentation as a tool for finding answers to policy questions. Experiments may be expensive compared with traditional forms of social research, but even the costs of major experiments are small compared with the costs of social policies that do not work or that might have been significantly more effective if experimental results had been available. I fully concur with Juster's position: economists and others who seek to improve the basis for policy formulation should devote far more attention and resources to data collection and measurement than they have in the past. Ever fancier statistical manipulation of inadequate data from traditional sources is unlikely to improve knowledge about policy choices significantly. The effort required to design and carry out both surveys and experiments may be painful and unglamorous, but it is the *sine qua non* of more rational and informed policy formulation.

The problem, of course, is not to choose between surveys and experiments, but to find the mix that will tell us the most. A reasonable strategy would seem to me to involve: (1) greatly increased resources for continuing panel surveys of many subpopulations whose behavior is of policy importance; (2) increased reliance on Type 1 experiments, often using part of a panel sample as the experimental or control group; (3) increased efforts to learn how to do Type 3 experiments with alternative methods of service delivery; and (4) very occasional use of Type 2 and 4 "experiments," mainly to develop or demonstrate a new policy and monitor results rather than as rigorous experiments to yield policy parameters.

REFERENCES

E. Gramlich and P. Koshel, *Social Experiments in Education: The Case of Performance Contracting,* Brookings Instit., Washington forthcoming.

F. T. Juster, "The Use of Surveys for Policy Research," *Amer. Econ. Rev., Proc.,* May 1974, *64.*

I. Lowry, ed., *Housing Assistance Supply Experiment, General Design Report: First Draft,* Rand Corp., Santa Monica 1973.

J. Newhouse, *A Design for a Health Insurance Experiment,* Rand Corp., Santa Monica, Nov. 1972.

G. H. Orcutt, "Data Needs for Computer Simulation of Large-Scale Social Systems," *Proc. of the Conf. on Computer Methods in the Analysis of Large-Scale Social Systems,* Joint Center for Urban Studies, Mass. Instit. of Tech. and Harvard Univ., Cambridge, Mass. 1964.

A. Rivlin and M. Timpane, eds., *Legal and Ethical Issues in Social Experimentation,* Brookings Instit., Washington forthcoming-a.

——— and ———, *Planned Variation in Education,* Brookings Instit., Washington forthcoming-b.

Abt Associates, *Housing Allowance Demand Experiment, Evaluation Design,* Cambridge, Mass. 1973.

Institute for Research on Poverty, *Final Report of the Graduated Work Incentives Experiment,* Madison, Dec. 1973.

Rand Corporation, *Technical Analysis Plan for Evaluation of the OEO Elementary Education Voucher Demonstration: Technical Dissertation,* Santa Monica, Feb. 1972.

The Use of Surveys for Policy Research

By F. THOMAS JUSTER*

How can survey research play a more effective role in the design, implementation and evaluation of public policy? To answer this question it is useful to look backward as well as forward. Typically, we tend to learn a good deal from our failures: obvious failures bring on both recriminations and postmortems, and in the process of finding someone to blame as well as trying to find out what went wrong we often do learn something about the latter. But we can also learn from our successes, at least if we can distinguish policy successes that are due to wisdom from those that are due to luck. It is not easy to base this distinction on casual observation, since policies that were adopted long ago and proved ineffective may no longer be around to be observed, while successes that resulted from luck would still be with us and are apt to be part of any current sample. First, however, let me look at the process that seems to be required for successful public policy formulation.

In principle, cost-effective public policies are the result of a formidable undertaking imposing extraordinarily heavy demands on the research community. Research inputs must answer the following questions:

1. What is the problem? What is it that we don't like and would like to change? For example, do we have too much or too little unemployment, too much or too little price inflation, unwarranted differentials in earnings for blacks and women, too much crime in the streets or alienation among the young?

2. What are the quantitative dimensions and structure of the problem? Which groups are affected both directly and indirectly by it; are its effects important or trivial, temporary or permanent; is it easing or getting worse? For example, is high unemployment due mainly to increased job search and other essential frictions, to deficient demand or to structural malfunction? Are rising welfare costs due to more eligible people applying or to more people becoming eligible? Are high school and college dropouts people who were pulled out by opportunities in the job market or pushed out by an ineffective and stultifying educational experience? Are the effects of price inflation perceived to be much worse than they are objectively? Who bears the distributional burden of price inflation?

3. What are the desired outcomes? That is, what do we want the policy to accomplish, by way of both short- and long-term impacts? I judge that it is only possible in principle to have policy changes in which some are better off and no one is worse off: in real life, all policies make some people worse off and some better off, and the issue is whether the resulting aggregate gains and distributional effects represent a net improvement.

4. How are changes in policy variables related to changes in situations or outcomes, and how are changes in situations related to changes in behavior, both in the short and long run? For example, the effect of subsidized health insurance depends on whether it increases the supply of health care, on the elasticity of demand for health services, and on the coverage of reimbursable expenses. Optimal policy to ameliorate the present energy crisis de-

---

* Program Director, Survey Research Center, and Professor, Department of Economics, The University of Michigan.

pends on demand elasticities, which can be expected to be much larger in the long run than in the short run, on the shock effect of a crisis in inducing people to accept temporary inconvenience, which can be expected to be important in the short run but not in the long run, and on public perceptions of the equity and fairness with which any system of controls is administered.

To develop a model for use in analyzing such questions, one must meet a number of imposing analytical and empirical requirements. The model must be specified fully and the relevant parameters must be estimated accurately. The model must contain at least one variable that can be manipulated by policy, it must include all the variables that have an important influence on behavior (which is different from having a "statistically significant" influence), and it must include both impact and long-term equilibrium effects. The latter are especially important, since in many cases the outcome desired by policy represents a quite long-run situational or behavioral phenomenon, and the identification of impact effects is useful only insofar as they constitute reliable predictors of these long-term or equilibrium effects.

5. What are the costs and benefits of alternative policies? All policies entail some costs; obviously we want to adopt policies that show the most benefits for a given cost level. But all the benefits and costs are relevant, not just those that are easy to measure or have already been measured. In particular, uncertainty warrants explicit consideration on two counts. First, uncertainty about the impact of any given policy constitutes a cost. Other things equal, an expected benefit with a given variance is less desirable than the same expected benefit with a larger variance. Second, since some uncertainty about outcomes is inevitable, provisions for monitoring the impact of a policy change is essential to prevent unintended effects from creating outcomes with high negative values.

I. Models and Public Policy

If we had to answer the questions spelled out above before we could formulate policies it seems clear that there wouldn't be any public economic policies. Simply to describe the formal requirements for ideally effective policymaking is to suggest that we are not now, and never have been, in a position to meet those requirements. But that is saying that policy must be framed with some degree of uncertainty about outcomes, which is hardly a novel observation. In the policy arena, the counsel of perfection is itself clearly a suboptimal policy. Whether or not the research community participates, policies are going to be formulated and adopted; the only issue is whether the research community provides some, admittedly imperfect, input or none at all.

The results of research inputs into the formulation of policies are mixed. While we have certainly had policy failures that can be attributed at least in part to inadequate research inputs, research seems also to have contributed to some important policy successes.

It is instructive to look at some of the failures and successes, since I read them as containing a clear-cut lesson for the future. Although my sample of successes and failures is basically impressionistic and a function of my own research interests, I do not think that other lists would convey a different message. Among the successes, ones that come to mind are: (1) Post-World War II policies to avoid major recession; (2) income-maintenance policies such as social security and unemployment compensation; (3) policies to provide mass public education; and (4) subsidization of home and automobile ownership.

With the possible exception of the last item, these represent major policy decisions designed to achieve certain objectives undertaken with at least some support from the research community. The incentive systems which have encouraged the growth of private homes (loan guarantees, tax advantages) and automobiles (the system of free highways) may or may not have been desirable economic and social policy, but the objectives of each policy, and to a considerable extent the benefits alleged by the proponents, have in fact been realized. The question, however, is whether these cases fall into the category of outcomes resulting from chance rather than from analysis.

Among the failures, my sample would include: (1) attempts to fine-tune the macroeconomy; (2) manpower training and retraining programs; (3) poverty and income redistribution programs; (4) attempts to increase the cost-effectiveness of public services like health and education; and (5) the urban renewal program.

It is useful to try to identify the distinguishing characteristics of policy successes and policy failures. In general, successful policies tend to be those addressed to basic and highly visible quantitative deficiency which could be largely if not entirely ameliorated either by direct grants of public funds to clearly identified groups or for clearly identified purposes, or by quite broad monetary and fiscal policies. In contrast, policies which attempt to fine-tune the circumstances or opportunities of particular groups, or the way existing policies were implemented, tend to be less successful.[1] In brief, we can simply say that policies requiring an axe have been reasonably successful; those requiring a scalpel have not.

For example, contrast our successful avoidance of major recessions with our inability to fine-tune, and our successful social security program with our relatively ineffective poverty program. Our present state of knowledge and the sophistication of macromodels make it hard not to avoid major recessions: we know enough about what is currently going on in the economy, via an effective monitoring system, and we know enough about how to change situations in relatively crude ways, via tax changes, expenditure changes, and monetary changes, so that it is probably the unanimous and correct judgment of economists that major recessions could only be due to egregious mistakes that could not be reversed quickly enough. While behavioral responses cannot be predicted accurately, the range of uncertainty is small enough to avoid gross errors; in addition, policies can be modified as behavioral responses become a datum. On the other hand, despite sophisticated macromodels, we do not know enough about lag structures, how expectations affect behavior and how they are formed, how households and firms respond to small and perhaps transitory changes in financial magnitudes of various sorts, etc., to be able to avoid the kind of excesses that periodically produce economic slack and unemployment of modest dimensions, nor to avoid periodic bouts of not so modest inflation. All our attempts to fine-tune founder on the rock of inadequate behavioral specifications in the model, and there is probably no cure for it at the macro level.

Similarly, the recognition that older

---

[1] Still another way to describe this difference is to say that the dependent variable in successful policy cases could be explained satisfactorily either by implicit (judgmental) models or by relatively crude macromodels, while the dependent variable in policy failures could be explained only by models with a high degree of accurate specification capable of predicting the impact of marginal changes whose effect may be substantially delayed in time and may not show up at all in the short run.

people are quite apt to be inadequately financed and thus to lead bitter and unrewarding lives, as well as create serious problems for their families, could be handled simply by providing some basic income support to most of the aged. The risk of having that income support spill over into undesirable labor force effects is largely absent because the aged are basically out of the labor force to begin with. But in the poverty program, just providing income to poor people runs up against the potential cost of providing strong disincentives for poor people to work and thus reducing aggregate output, while trying to structure a range of poverty programs designed to ameliorate the circumstances of particular groups with special characteristics, different incentives, different needs, living in different parts of the country with local poverty programs that provided sharply differential support levels, has proven to be, so far, a relatively intractable problem.

Some would argue that we know how to solve the poverty problem but simply lack the political will to do it. I would tend to disagree, since I think the disincentive effects of income support are not to be taken lightly, particularly in terms of their potentially deleterious effects in the long run when everyone has fully adjusted to their opportunities and the relative risks and rewards of various choices.

II. Alternatives for Current Policy Issues

We have devised policies in certain areas where the benefits of change were large and the costs small, and where a reasonably appropriate policy could be specified with a relatively modest input from the research community. But the policy issues that now confront us are more difficult, the desired outcomes more complex, the specification problems more important, and the risks of adopting policies with negative net benefits higher. Thus, the kind of research input that was good enough in the past will not be in the future.

In this new environment, we can approach the problem of specifying appropriate behavior relationships for input into policy formulation in two general ways. (1) We can use whatever natural experiments are cast up by the economic and social system and spend a lot of resources in creating good measurements of micro phenomena and on devising appropriate micromodels in which these data inputs can be used. In general terms, I focus on this course of action. (2) Where good natural experiments are not available, or where the empirical specification of the model is uncertain because we don't fully understand how the data were generated and have difficulty in choosing between alternative models, we can use even more resources in an attempt to generate experimental data where we know exactly how the data were generated (See Alice Rivlin.)

Note that both alternatives imply that the solution to many current policy issues depends on the application of more resources to the measurement of social phenomena: for natural experiments, the argument is that the basic shortcoming is the deficiency of sufficiently accurate and relevant microdata; for controlled experiments, the argument is that more accurate measurement and modeling cannot resolve the uncertainties, and it is necessary to use even more resources in order to generate the observations needed to test hypotheses that bear importantly on policy issues.

At bottom, that is all the New Jersey and other negative income tax experiments do. We do not have any natural experiments in which one could quantify the labor supply response to an income guarantee that also contained a marginal tax rate on earnings substantially under 100 percent. If the labor supply response is in dispute, there is no other way to resolve the debate except by generating observations where a set of income guarantees and marginal tax rates is created via experiment.

III. Policy Issues and Research Strategies

Social scientists are by no means in agreement about the practical importance of research for policy. The majority view probably is that the substantive issues discussed in this paper are basically irrelevant, and that the real problem is that political leaders are simply unwilling to ask society to pay the cost of what many believe to be optimal policies. There aren't enough votes for those policies because some people stand to lose from them and many others are uncertain about whether they will gain or lose; thus policies that are socially desirable—or are seen by some as desirable—fail to be adopted. An alternative view is that we don't really know how to solve these problems or what the benefit-cost ratios for different policies really are, and that political leaders know this and, hence, are unwilling to vote programs whose outcome is unknown or highly uncertain and whose costs are visible and quite high. I would argue for the latter view, although, even if the first is correct, the uncertainties implied by the second provide adequate ammunition in the hands of those who are unenthusiastic about a program to start with.

IV. A Note on Research Strategy

In economics and the social sciences generally, the basic research strategy has been heavy reliance on either implicit models or macromodels that make relatively modest demands on data and can often be satisfied with data obtained for administrative purposes by federal statistical agencies. That's clearly a sensible way to start solving "gross" or macro-level problems and the results of applying existing models to aggregate or other (often conceptually inappropriate) data are sufficiently precise to be useful. But for micro-level problems one needs both sharper tools of analysis and data capable of testing refined and sensitive models. The application of sharper and more sophisticated tools of analysis to a relatively crude aggregate data base will produce better insights for a while, but eventually this becomes an exercise in futility.

Thus my conclusion is that the success of analysis and prediction which is essential for the specification of accurate policy outcomes depends largely on our ability to generate and use relevant micro observations with the required degree of precision. But social scientists and economists generally have never adopted that as a sensible research strategy, for reasons that are clear enough.

1. There isn't much professional payoff—the highest professional payoffs accrue to model developers, not to measurers of empirical phenomena.
2. That strategy requires financial support well in excess of prevailing levels. For example, a recent proposal that I designed with this focus in mind came to 60 percent of the total annual budget of the Economics Branch of the Social Science Division at National Science Foundation. Nobody, including me, would really argue that a single project warrants that large a fraction of research support for the profession generally.
3. The strategy requires a commitment to a longer gestation period than research funding organizations typically can, or feel they can, afford to allow.
4. It requires a broad professional consensus about the importance of measurement that is largely lacking at present, although one can find evidence that more professionals are tending to see the problem in this light.[2]

---

[2] See the comments by Kenneth Boulding and by Irwin Friend; see also Harry Johnson who suggests the risks of ill-judged policy but does not suggest a solution.

It is informative to note some contrasts between the way in which different scientific disciplines function in regard to the relative emphasis in the research process on theory (modeling) and measurement.[3] Let us take as given the hypothesis that the ability to predict unknown outcomes is the true test of the scientific validity of any discipline and also a valid test of the ability of that discipline to contribute to the solution of policy issues. Ability to predict unknown outcomes differs markedly among disciplines; in general, it tends to be pretty good in the physical sciences and pretty poor in the social sciences.

The most striking difference in research methodology between sciences is in their inherent capacity to generate experimental observations and in their allocation of resources between theoretical work and the measurement of empirical phenomena. In the physical sciences, the generation of experimental observations is a way of life and the accepted path to knowledge, and the allocation of resources between theory and measurement reflects that way of life. In the social sciences, the situation is exactly reversed.

Part of the explanation, of course, is that social scientists are precluded from making the same kind of experimental observations as physical scientists, since we have to intrude on the economic and social system (on people) in order to make the observations while scientists aren't hampered with that issue. Electrons don't care if they're being observed, or at least we don't think they do, and nonhuman organic substances and nonhuman life do not complain about invasion of privacy. Hence, the basic research strategy in the social sciences is to measure naturally occurring phenomena and hope that we can disentangle the causal relationships via statistical techniques.

V. The Role of Survey Research

The thesis of this paper is that the basic reason for policy failures of the sort we have experienced is that resources for research in economics and other social sciences have been misallocated and that our policy research will not begin to yield the same kind of fruits as are commonplace in the physical sciences until we pay more attention to measurement.

Surveys can meet policy needs in five different ways. First, they can provide a description of the basic structure of an alleged or potential policy problem. For example, how much poverty and income inequality is due to conscious choices about leisure or the leisure component of work, training and future income versus no training and current income, consuming now or consuming later, the acceptance of regular or intermittent employment as a preference, and so on? And how much observed inequality is a statistical artifact of life-cycle differences that reflect differences in needs and investment patterns but not differences in opportunities? To the extent that income inequality reflects either conscious choices about work and leisure in the broadest sense, or to the extent that it simply reflects life-cycle patterns, it would not generally be regarded as an important concern. But to the extent that it reflects differences in opportunity sets, it is a matter of concern.

To turn to still another area, what is the distribution of energy use between market-related activities (commuting, etc.), home production (shopping, chauffeuring, heating), and recreation (leisure) activities? What is the distribution of potential rates of return to energy-saving investments? How much do people know about the potential rate of return to, for example, home insulation, replacement of pilot lights with mechanical starters, etc., particularly when energy prices are higher than they have been historically, and how quickly do they react to new circumstances? How can reactions and adaptations be speeded up? I do not see how we can react sensibly to the present energy crisis without access to this kind of simple descriptive information, and at the present time we apparently do not have it.

---

[3] This section draws heavily on my 1970 paper.

Of course, there are several ways to get relatively crude descriptive information about the dimensions of a problem; a partial substitute for surveys might be administrative records, for example, social security files. But to the extent that expectations, preferences, and perceptions must be measured there are no alternatives to the survey route.

Second, and possibly most important, surveys can provide the data inputs needed for better estimation and specification of the behavioral models that are essential for accurate policy research. The simplest contribution that surveys can make is to provide measures of variables that are important in macromodels but cannot be measured accurately with standard data collection techniques. For example, there appears to be no wholly satisfactory way to measure expectations from existing data on actual change; the reason is that expectations are not a simple function of past experience. The most useful measures of expectations are obtained via surveys, as a number of recent studies of price expectations have shown: both in models that explain wage behavior and in those that explain savings, expected price changes appear to play an important role.[4] And the variance in expected price among the population seems to be an extremely important variable in the personal savings function (Juster 1973). To the extent that expectations are a significant ingredient in macromodels, surveys are essential to provide the measurement.

Surveys can also provide the data base for the micromodel counterparts of macromodels. Micromodels can be substantially richer because they can contain many more variables (because the microdata admit of more degrees of freedom) and can be disaggregated so that, where appropriate, different functional relationships for different population subgroups may be specified. Particularly when it comes to specifying dynamic processes, the use of micromodels with a longitudinal data base seems the only feasible strategy for obtaining correct estimates.

Survey data may also serve as inputs for simulation models. It can be argued that the complexity of the macro universe is such that the only feasible way to estimate macro effects is to create a complex micromodel which allows a wide range of differential effects of different variables on behavior, depending on the particular characteristics of the unit, and then simulates the macro universe by aggregating micro predictions. Basically, the argument is that the appropriate macromodel is a complex aggregation of a wide range of appropriate micromodels, and the problem of specifying the correct macromodel cannot be handled analytically but must be done empirically by simulation. In such a model system, it is possible to identify the parameters to which the model is most sensitive and then use survey or other data to estimate parameters that seem to be most crucial and least well defined.

In the process of building more accurate microdata bases either for standard micromodels or for simulation models, interaction between model builders and measurers is essential. A number of massive and expensive data collections have remained unused—often because an administrative agency required the data to answer a particular question and designed the operation to do only that. Examples abound, although they may have been more characteristic of earlier periods than recent ones.

---

[4] See George de Menil and Surjit Bhalla, Juster and Paul Wachtel, and Lester D. Taylor.

Interaction between model builders and data gatherers is important because all model builders feed on empirically observed phenomena in one way or other, and the richer the observed phenomena that go into the construction of models, the richer the model. The standard conception of how insights are gained in the real world is that casual observation generates a crude model, the performance of the model suggests new insights and the need for new observations, the new observations generate a revised model which in turn generates a need for new observations, and eventually that process produces a reasonable approximation to scientific truth. In addition, interaction between model builders and measurers may be the best way to induce technological change in measurement methods, which tend to be pretty backward in the social sciences.

A third area in which survey data can contribute to policy issues is in the measurement of preferences for different outcomes. For many policy issues, it is important to specify the preferred set of outcomes in order to determine whether a specified pattern of changes would represent a gain or a loss, or whether one of two alternatives involving the same cost represents the largest gain or the smallest loss. An immediate case in point is the best method of meeting the present energy shortage, where in the short run we are clearly going to have to curtail consumption rather than expand supply. Of the various methods by which consumption might be curtailed, which is perceived to be the least damaging? To the extent that the price mechanism is used as a rationing device, that question can be ignored or avoided, since the market will solve the problem and we don't have to know in advance how it will do so—except that it may make a difference whether equilibrium is attainable with 60-cents-per-gallon gasoline or only with $2-per-gallon gasoline. But to the extent that other methods are used, and that seems probable given political and public attitudes, it is not at all obvious which kind of curtailment will do the least damage to economic and social welfare.

Income distribution policy provides another current illustration. Does the acceptance of a given amount of redistribution depend on the mechanism by which it is accomplished (money, income-in-kind), as Tobin has argued? (See Tobin.) If so, does that represent a stable preference function or one that can be modified by educating people to the inequities and inefficiencies of income-in-kind redistribution? Do redistribution programs really require the appearance of doing something for everybody, so that the gross size of the program is some multiple of its net effect?

Fourth, surveys may not only permit the measurement of public preferences in cases where the issue is simply what do people really want, but they may also provide information which could be useful for those seeking to educate the public about the costs and benefits of alternative policies. We know very little about how people shape their preferences, although informational inputs must play some role in helping people to decide whether, for example, they prefer the negative income tax to the present welfare system. But it will often be true that popular perceptions of alternatives are biased, in that preferences are unstable and capable of being modified or reversed by an input of accurate information. The issue here is not one of propaganda or salesmanship; rather it is whether current preferences are the same as those that would exist if everyone had easy access to the best current information concerning a particular policy issue. Survey research can detect where people get information from, whether certain types of informational inputs are correctly perceived, and even whether preferences would alter if informational inputs were perceived differently. In sum, survey methods can be used to explicate public preferences for goods and services that must be purchased communally and cannot be bought by an individual as he chooses; they can be used to judge whether preferences are based on misinformation or on misconception; and possibly they could be used to determine the best way to increase the effective level of information.

Fifth, surveys offer a means of monitoring the economic and social system so as to get an "early warning" of emerging problems and to provide a means of delimiting the scope of existing and emerging problems as quickly as possible. It would be unwise to assume that we can predict what kinds of policy issues will emerge in the future, and the best we can hope for is to recognize issues early rather than late. The issues often arise from the delayed effects of past policies or of interactions between past policies and the dynamics of economic and social change. To the extent that issues arise from combinations of events that we do not fully understand, which is the usual case, we have no means of predicting them in advance from existing or prospective models, at least not for a very long period of time. Thus the only sensible approach is to recognize our inadequacies in this regard and to count on continual monitoring to show up potential problems at an early stage.

For example, it is hard to believe that systematic monitoring of the urban renewal or welfare programs would not have turned up increasingly strong evidence that these programs were not producing the positive outputs that were hoped for and were, in addition, having some important negative effects. In the financial research area, we get very little help from aggregate data in understanding the present behavior of the securities market, the real estate market, or the foreign exchange and overseas investment market. Knowing aggregate or hypothetical returns does not get us very far in knowing the distribution of actual returns among classes of savers and investors, in knowing about perceived opportunities and *ex ante* returns, or in predicting the impact of aggregate policies on saving and investment flows, portfolio shifts, and wealth distributions.

This fifth use of survey research suggests that we should consider the potential usefulness of a vehicle whose basic function is to monitor economic and social change, to test hypotheses about behavior, to explicate the dynamics of behavior change, and to provide a resource for policy formulation. What I have in mind, simply stated, is that we need to consider the creation of a "permanent" panel which would represent a microcosm of the economic and social system and for which we would accumulate information of the accuracy and type that would be needed for a wide but unpredictable range of policy issues in the future. Such a proposal would have been unrealistic some years back, simply because the problems of managing large data bases would have made such an attempt an expensive fiasco. But computer and data-handling technologies have advanced to the point where even extremely large data bases can be handled comfortably and relatively inexpensively. Experience with repeated interviews of a panel suggests that problems of panel mortality, bias, and mobility are manageable. Hence, the relevant question is: How much would such a panel cost and how much would it be worth? We know something about the cost—several million dollars a year for an indefinite period of time. We know less about the potential benefits except that they are highly uncertain. But if more effective research inputs can avoid the loss involved in even one poorly conceived national policy each decade, or promote the gains involved in one well-designed policy, the benefits will be a large multiple of the costs—substantial though the latter may be. As an investment strategy, it looks to be an attractive gamble.

REFERENCES

K. Boulding, "Equality and Conflict," *The Annals*, Sept. 1973, *409*, 1–8.

G. de Menil and S. Bhalla, "Direct Measurement of Popular Price Expectations," Econometric Res. Prog., Princeton Univ., res. memo. no. 149, July 1973.

I. Friend, "Mythodology in Finance," work. pap. no. 1–73, Rodney L. White Center for Fin. Res., Univ. of Pennsylvania, The Wharton School.

H. G. Johnson, "Some Microeconomic Reflections on Income and Wealth Inequalities," *The Annals*, Sept. 1973, *409*, 53–60.

F. T. Juster, "Microdata, Economic Research, and the Production of Economic Knowledge," *Amer. Econ. Rev., Proc.*, May 1970, *60*, 138–148.

———, "Savings Behavior, Uncertainty, and Price Expectations," prepared for the Conf. on the Econ. Outlook, Ann Arbor, Nov. 1973, unpublished.

——— and **P. Wachtel**, "Inflation and the Consumer," *Brookings Papers on Econ. Activity*, 1972:1, 71–121.

A. Rivlin, "How Can Experiments Be More Useful?" *Amer. Econ. Rev., Proc.*, May 1974, *64*.

L. D. Taylor, "Saving of U.S. Households: Evidence from the Quarterly Flow-of-Funds," Dept. of Econ., Univ. of Michigan, unpublished.

J. Tobin, "On Limiting the Domain of Inequality," *J. of Law and Econ.*, Oct. 1970, *13*, 263–77.

INTERCITY PASSENGER TRANSPORTATION

Airline Capacity Limitation Controls: Public Vice or Public Virtue?

By GEORGE EADS*

Are airline capacity limitation controls in the public interest? The U.S. Civil Aeronautics Board (*CAB*) itself once labeled them "inconsistent with the general competitive norm established by the Federal Aviation Act and with the antitrust laws" and has refused to permit the carriers to discuss such arrangements "except in highly unusual and limited circumstances" (*CAB* order 71–8–91, p. 3). During the summer of 1971 the Board felt that these conditions were met in four major transcontinental markets and reluctantly approved a capacity agreement submitted by the three carriers involved.[1] Now two and one-half years after these agreements were approved, the views of the Board as to the appropriateness of capacity limitation agreements clearly are changing.

In a concurring opinion attached to a Board order extending the capacity agreements past their originally scheduled expiration date, the current *CAB* chairman, Robert D. Timm, wrote:

> . . . in the context of the highly competitive nature of the U.S. air transportation industry, excessive capacity is not a problem which can be effectively dealt with on a unilateral basis. The capacity agreement has proven to be an effective tool for eliminating destructive excessive capacity competition. The short- and long-term benefits of extended capacity agreements are many and quite meaningful to the public and the industry. For these reasons I would make the capacity agreement available for industry-wide use on a voluntary basis, subject to Board review and approval, providing the carriers rendered an accounting for the capacity which is freed by such an agreement. [CAB order 72–11–6, Timm concurrence, p. 14]

Subsequently, the Board approved discussions aimed at extending the existing agreements for an even longer period and encouraged carriers to suggest other markets for which such agreements might be made (*CAB* order 73–4–98). The carriers responded enthusiastically.[2]

According to their proponents, capacity limitation agreements are desirable for one of two reasons: without them, competition among carriers might be either "destructive" (i.e., lead to an equilibrium in which all carriers operate at a loss) or "wasteful" (i.e., lead to an equilibrium in which carriers make a profit but have very high levels of excess capacity). Is either of these arguments valid?

---

* Associate professor of economics, George Washington University. I would like to thank Robert Goldfarb, Mark R. Killingsworth and James C. Miller, III, for helpful comments.

[1] *CAB* order 71–8–91. The markets involved were New York/Newark-Los Angeles, New York/Newark-San Francisco, Washington/Baltimore-Los Angeles, and Chicago-San Francisco. Later a similar agreement was approved in the New York-San Juan market (*CAB* order 72–6–70).

[2] As of the time this paper is being written (late December 1973) capacity limitation agreements are either in force, are awaiting approval by the *CAB*, or are under discussion by the carriers in over fifty markets (*Aviation Week*, p. 26).

I. Is Airline Competition "Destructive"?

The argument that service competition among airlines is "destructive" usually goes as follows: The carrier offering the largest share of capacity (usually measured by flights or seats) in a given market will capture a disproportionately dominant share of the market's traffic. Knowledge of this phenomenon spurs all carriers to add capacity in the hope of obtaining market dominance, and as all attempt to follow this strategy, more capacity is added than the market can absorb, driving down load factors and bringing losses to all parties. Any carrier trying to opt out of this capacity race will suffer a disproportionate loss of traffic to his rivals and will also suffer financial losses, though perhaps on a smaller scale.

This model was first formally proposed by W. E. Fruhan, Jr., who cast the airline scheduling decision explicitly in terms of a Prisoners' Dilemma game and illustrated by means of hypothetical examples how such a loss equilibrium might be achieved in a two-carrier market.

If maximization of market share was itself the major goal of the airlines, such a model might explain capacity competition, provided the hypothesized nonlinear relation between capacity share and market share were found to exist. However, the airlines are presumed to be interested in profits, not market share per se. Consequently, it is important to consider the incentive to add or reduce capacity in a given market in terms of the change in profit that is likely to result.

Following G. Renard (p. 34), the hypothesis of a nonlinear relationship between capacity and market share can be written as

$$MS_i = \frac{F_i^a}{\sum_{j=1}^{n} F_j^a}$$

where MS_i is the market share of the ith competitor, F_i is a measure of the capacity (either seats or flights) offered by the ith competitor, and a is a parameter, $a>1$ by assumption. As Renard has shown, if $a>1$, this expression gives rise to an S-shaped relationship between capacity and market share.

Even though a loss equilibrium is possible under such a model, both J. V. Yance (1970) and G. W. Douglas and James C. Miller III (forthcoming 1974) have shown that contrary to the impression conveyed by Fruhan, this result is by no means inevitable even if $a>1$. Specifically, a loss equilibrium will occur with this model if and only if $a(n-1)+e<n$, where n is the number of competitors in the market; and e is the elasticity of traffic in the market with respect to changes in the total number of one-way daily flights offered by all competitors, F.

Renard (p. 37) estimates that $a=1.45$. Using more recent data, Douglas and Miller (forthcoming 1974) estimate $a=1.26$. Yance reports findings by others that $e=1.46/F$ (1970, p. 4). Thus, for a two-carrier market, a loss equilibrium is possible only in markets where the total number of one-way flights per day is less than two (using the Douglas-Miller estimates) or three (using Renard's). In three-carrier markets the critical number of one-way daily flights is four and fifteen, respectively.

Some may already have noticed that as the number of competitors increases, loss equilibria become possible in all markets regardless of the frequency of service offered. This is true for markets with four or more competitors if Renard's estimate of a is used or for markets with five or more competitors if the Douglas-Miller estimate is employed. This result should not be given great weight for two reasons. First, there is some evidence that a may decline with the number of carriers

(Renard, p. 39).[3] Second, although more than three carriers are often certificated to serve a market (Renard reports twenty-five six-carrier markets in his 1967 sample, p. 39), there are seldom more than two or three carriers whose actions would be expected to exert a significant influence on carrier-scheduling decisions. In certain cases, carriers are prohibited from offering turn-around service in a given market, so must base their scheduling decisions on considerations other than the expected action of rivals in the market in question. In other cases, carriers are certificated in a market only on the condition that they serve an undesirable airport. The distinction is usually made, therefore, between the number of carriers offering service in a given market and the number of "effective competitors," defined as carriers possessing at least a 10 percent market share. If we can assume that only the actions of "effective competitors" normally would have an influence in scheduling decisions, then it is quite apparent that the two-carrier and three-carrier results are the appropriate ones to look at, for in 1970 only three of the top 135 city-pair markets had more than three effective competitors. Four were effectively monopolies, 90 had two effective competitors, and 38 had three (Eads 1971, p. 5). Furthermore, there is evidence that most city-pairs not among the top 135 city-pairs were served by no more than two effective competitors. The average number of effective competitors serving the top fifteen city-pairs ranked by passenger miles was 2.80; the average number serving the bottom fifteen city-pairs was 2.07.

Consequently, it does not seem unreasonable to conclude that loss equilibria are by no means inevitable even if one accepts Fruhan's model, provided the *CAB* has added a third competitor only in markets with a level of traffic sufficient to support a reasonable number of one-way daily flights. And, as we have just seen, this usually has been the case.

There are even stronger reasons to believe that airline competition on the basis of service rivalry is not "destructive." While Fruhan's treatment of the airline scheduling process as a Prisoner's Dilemma game is interesting, the mere fact that it can be thus depicted does not mean that the model accurately describes the process.

The Prisoner's Dilemma reaches the outcome it does because the two prisoners are unable to communicate with each other (R. D. Luce and H. Raiffa, p. 95). Each prisoner knows the consequences both of talking and of keeping silent. If communication were possible, neither would confess and both would escape heavy prosecution. However, communication being impossible, each prisoner confesses, since that action minimizes the expected sentence. This produces a result that neither desires.

This reveals the weakness of the game as a model of airline scheduling. The payoff matrix proposed by Fruhan obtains only if each firm can be assumed to believe that it can add capacity without its rivals reacting. This would be a reasonable assumption only if adding capacity was something that a competitor could do secretly or if capacity additions, once discovered, could not be easily matched. But to paraphrase George Stigler (p. 44), no airline has yet been able to devise a way to advertise schedule changes which brings them to the attention of numerous customers but not to that of any rival.

We would agree with Renard (p. 45) that a much more realistic attitude for an airline to take is that additions to schedules are likely to be matched while reductions may not be. In adopting this assumption we reduce substantially the probability that a loss equilibrium will occur regardless of the number of competitors, "effective" or otherwise, in a market. This is because a carrier contemplating the addition of a flight will not be able to anticipate winning passengers away from its rivals but instead will gain only a proportional share of the additional traffic generated by the increase in total frequency. As A. S. DeVany has shown (pp. 4–5), in equilibrium the elasticity of traffic with respect to frequency will always be less than unity. In long-haul markets particularly, this elasticity should be quite low once even a few one-way flights are offered per day.[4]

---

[3] We are skeptical of these results for two reasons. First Renard's division of his sample by number of competitors is subject to the criticism advanced immediately following in the text of the paper. Even Renard appears to admit this as a possible weakness (p. 37). Second, Renard's estimates by number of competitors generally are not significantly different from each other.

It is, therefore, quite unlikely that carriers who take this view as to the likely reaction of rivals will be misled by false profit expectations to add so much capacity in any given market that a loss equilibrium ensues. The more probable result is the one suggested by Douglas and Miller (forthcoming, ch. 4) and by Yance (1972, p. 55)—that capacity will be added only up to the point where the last flight just breaks even.

A problem might arise if, because of mistaken perceptions by one or more carriers as to rivals' possible responses,[5] a decline in demand, or a deliberate policy of overscheduling by one of the competitors, the carriers in a market found themselves operating at a loss. The model we have proposed suggests that carriers might find it difficult to extricate themselves relying solely on unilateral action. Each might be reluctant to initiate a "strategic retreat," fearing that its rivals would not follow.

Again, we must be careful not to confuse maximization of market share with maximization of profits. The above dilemma would be a serious one only if the carrier were unable to reduce its costs due to its cut in schedules by more than it reduced its revenues. The fact that all flights are not of equal attractiveness to passengers suggests that it should be possible in most cases to raise average load factors by unilateral schedule reductions. In the absence of peak-load pricing, demand tends to cluster at certain times of the day. All carriers serving a market seek to cover these periods of peak demand first and then add flights at less and less attractive times. Consequently, if the carriers serving a market found themselves in a position where all were losing money, each would be likely to be able to identify flights which could be eliminated with a loss of relatively few passengers. Cutting these flights would serve not only to raise the carrier's average load factor in the market but would also serve as a signal to rivals of its willingness to engage in a "mutual withdrawal."

We strongly agree, therefore, with the position taken by the Board in its decision in Phase 6B of the Domestic Passenger Fare Investigation—that "*non*destructive" scheduling competition is both possible and routinely occurs in the airline industry, even in the midst of an industry downturn (*CAB* Order 71–4–54, p. 30). Yance's study (1972) of scheduling competition during the early 1960's when the jets were being introduced provides additional supporting evidence.

Although we do not believe it likely that the carriers serving a market will be misled by false market signals into adding capacity until all suffer losses, we do not rule out the possibility that a carrier might, *as a deliberate policy*, initiate a "scheduling war" in the full knowledge that it could lead to temporary losses both for it and for its rivals. Such a strategy might be engaged in for the purpose of disciplining a competitor or influencing *CAB* policy.[6] Cases of deliberate overscheduling are known to exist, and several are cited by Fruhan (pp. 131, 132, 167 fn. 15).

---

[4] Recall the results cited above that the elasticity of total traffic demand with respect to flight frequency equals $1.46/F$. Thus, if F is as low as five one-way flights per day, $e=0.29$.

[5] This appears most likely if one of the carriers is unfamiliar with the way the scheduling "game" is played. For example, during the late 1960's and early 1970's, local service carriers were allowed to compete for the first time in a significant number of markets on a head-to-head basis with trunklines (Eads, 1972).

This possibility of "predatory scheduling" provides a strong reason for the *CAB* to be wary of endorsing the principle of capacity limitation agreements—even if for only a temporary period. A firm contemplating undertaking such a policy would compare the benefits (in the form of increased future profits) against the costs (in the form of profits foregone during the duration of the "war"). If capacity limitation agreements are viewed by the Board as a valid solution to low load factors, "predatory scheduling" may be *encouraged* since the availability of an institutionalized arrangement of ending such "wars" should lower the cost of engaging in them.

II. Is Airline Competition "Wasteful"?

Another group of proponents of airline capacity limitation agreements argues not that airline competition is "destructive" but that it is "wasteful." These people do not quarrel with the proposition that a carrier can exercise unilateral control over its capacity, but point to the fact that, at least until the energy crisis, airlines on the average flew with almost half of their seats unsold and that the proportion of unsold seats had been rising throughout most of the postwar period (*CAB*, 1970, p. 26).

Any system that produces a nonstorable product and that faces a stochastic demand must operate with some level of "excess capacity" on average if the public is to receive adequate service (Douglas and Miller, forthcoming 1974). The relevant question, therefore, is not whether the airline industry operates with excess capacity, but whether it operates with "excessive" excess capacity. As I have argued elsewhere (1971), the answer to this question depends on the balance struck between the cost of providing excess capacity and the price air travelers are willing to pay for seat availability and other amenities. Reasonable estimates of the parameters involved indicate that the level of excess capacity currently being provided may indeed be excessive (Douglas and Miller, forthcoming 1974). However, capacity limitation agreements are not the answer. As L. J. White has demonstrated, such agreements are likely to produce an *insufficient* level of excess capacity. The cost to the public may be far from trivial. The U.S. Department of Transportation (*DOT*) estimates that based upon the cost-based fare, the allocable investment, and the existing yield, the carriers could earn as much as a 40 percent rate of return on investment in the New York-Los Angeles market under the terms of the current agreement. Looked at another way, each passenger flying this market may be paying as much as $52.56 (or 51 percent) more than the cost-based fare, given the level of service quality he actually is receiving (*DOT*, 1973, pp. 36–38).

The *DOT* correctly asserts that if load factors are too low in certain markets, the *CAB* can raise them by reducing fares. This forces up the break-even load factor, the proportion of seats that must be sold to cover costs. Carriers would have to cut back on schedules, raising the actual load factor. The resulting load factor might be the same as under a capacity limitation agreement (though for the reasons suggested above, it probably would be lower), but the public would be compensated by reduced fares for the decrease in service quality.

---

[6] One indicator that the *CAB* has employed in the past in deciding whether or not to add an additional competitor to a route is the load factors experienced by the existing competitors (Richard Caves, pp. 210–11).

The use of fare adjustments as opposed to capacity agreements to correct excessive capacity also avoids the likelihood that increased profits generated by capacity agreements may be partially or totally dissipated through other forms of cost-increasing, nonprice competition such as more sumptuous meals, wider and softer seats, more leg room, faster baggage and reservations service, more frequent changes of stewardess uniforms, more advertising of these amenities—in short, by anything that might serve to increase profits by attracting customers away from competitors given that both prices and capacity offered are fixed. Alternatively, as noted above, these profits could serve as a "war chest" to finance capacity competition in nonagreement markets with a view towards forcing agreements in these markets.

As long as price is fixed above marginal cost and some form of nonprice competition is possible, service inflation will tend to occur. The only alternatives to using fare reductions to control service inflation are for the Board to seek legislation permitting it to regulate every possible aspect of nonprice competition or to allow capacity agreements negotiated by the carriers to include agreements relating to all possible forms of nonprice competition. However, if the latter course were contemplated, the public would be better served by the decertification of all but a single carrier on any one route and the award of the franchise for the route by competitive bid, as Harold Demsetz once suggested with regard to public utilities.

Even more preferable would be to allow the carriers to experiment with a range of price-service combinations so as to be better able to meet public needs. For as Alfred E. Kahn (p. 216) has observed, it is not the service competition per se that is undesirable but the fact that air travelers have not had the opportunity to judge whether the improved service quality has been worth the higher fares.

This does not imply that total deregulation is the only solution. Under a proposal recently advanced by *DOT* (1972) and endorsed by the Department of Justice, Council of Economic Advisers, and American Airlines, carriers would be free without securing Board approval to raise or lower fares by up to 15 percent around a standard fare level set by the Board. Adoption of this "Zone of Reasonableness" approach would free the *CAB* from the virtually impossible task of determining exactly the price and service combinations the public wants and embodying them in the fare structure. Instead, the Board could adopt the fare structure it believed most closely reflected the public's relative preferences for low fares versus improved service; but if the Board misjudged slightly, unilateral fare and service adjustments by the carriers would compensate.

III. Conclusion

In a recent paper detailing their history, William Jordan concluded that airline capacity limitation agreements represent nothing more than an attempt by the *CAB* to close a loophole to the maximization of joint industry profits left open when the Civil Aeronautics Act of 1938 explicitly denied the Board the right to regulate scheduling and other aspects of service competition. While this interpretation may be open to some question, it does possess the virtue of not attempting to wrap airline capacity limitation controls in the cloak of "public interest." This is appropriate for as we have seen, the public is not well served by such agreements.

REFERENCES

R. E. Caves, *Air Transport and Its Regulators,* Cambridge 1972.

H. Demsetz, "Why Regulate Utilities?", *J. Law Econ.,* Apr. 1968, *11,* 55–65.

A. S. DeVany, "Some Aspects of Quality Competition in Regulated Industries: The Case of Airlines," unpublished, 1973.

G. W. Douglas and J. C. Miller, III, *Economic Regulation of Domestic Air Transport: Theory and Policy,* Brookings Institution, forthcoming.

——— and ———, "Quality Competition, Industry Equilibrium, and Efficiency in the Price-Constrained Airline Market," *Amer. Econ. Rev.,* forthcoming 1974.

G. C. Eads, "Competition in the Domestic Trunk Airline Industry: Excessive or Insufficient?", Brookings Institution, unpublished, Dec. 1971.

———, *The Local Service Airline Experiment,* Washington 1972.

W. E. Fruhan, Jr., *The Flight for Competitive Advantage: A Study of the United States Domestic Trunk Air Carriers,* Boston 1972.

W. A. Jordan, "Airline Capacity Agreements: Correcting A Regulatory Imperfection," *J. Air Law and Commerce,* forthcoming.

A. E. Kahn, *The Economics of Regulation: Principles and Institutions, Vol. II,* New York 1971.

R. L. Luce and H. Raiffa, *Games and Decisions,* New York 1957.

G. Renard, "Competition in Air Transportation: An Econometric Approach," unpublished M.S. thesis, MIT, September 1970.

G. J. Stigler, *The Organization of Industry,* Homewood 1968.

L. J. White, "Quality Variation When Prices Are Regulated," *Bell J. of Econ. and Mgt. Sci.,* Autumn 1972, *3,* 425–436.

J. V. Yance, "Nonprice Competition in Jet Aircraft Capacity," *J. Ind. Econ.,* Nov. 1972, *21,* 55–71.

———, "The Possibility of Loss-Producing Equilibria in Air Carrier Markets," unpublished, Dec. 1970.

Aviation Week, "Airlines Meet on Capacity Control As Fuel Saver," Oct. 23, 1973, *99,* 26.

U.S. Civil Aeronautics Board, *Handbook of Airline Statistics, 1969 ed.,* Washington 1970.

———, docket 21866-6B, order 71-4-54, decided Apr. 9, 1971.

———, docket 22908, order 70-11-35.

———, docket 22908, agreement 22496, order 71-8-91, adopted Aug. 17, 1971.

———, docket 22908, order 72-6-70.

———, docket, 22908, order 72-11-6, adopted Nov. 2, 1972.

———, docket 22908, order 73-4-98, adopted Apr. 24, 1973.

U.S. Department of Transportation, "Answer of the Department of Transportation to the Joint Application of American Airlines, Inc., Trans World Airlines, Inc., and United Air Lines, Inc., For Approval of Agreement," *CAB* docket 22908, June 15, 1973.

———, "Brief of the Department of Transportation to the Civil Aeronautics Board," *CAB* docket 21886-9, May 24, 1972.

Amtrak in Perspective: Where Goest the Pointless Arrow?

By EDWIN P. PATTON*

Amtrak, formally the National Railroad Passenger Corporation, has as its symbol a pointless arrow. Those who oppose the operation of intercity passenger trains in the 1970's consider the arrow appropriate for two reasons. First, the Corporation to date has failed to attain its goal of revitalizing rail travel and has required public subsidy in order to survive. Second, and a more basic criticism, there is no reason to subsidize a mode of travel that we as a society rendered virtually obsolete decades ago through dedicated public support of highway and air transportation.

This paper considers whether there is a point, or a purpose, for the pointless arrow. I describe the legislation that created Amtrak and analyze the Corporation's record in its potential markets. I then suggest criteria for evaluating the Corporation's future performance, and, finally, discuss political elements that may subvert this evaluation.

Amtrak assumed responsibility for most of the country's intercity passenger trains on May 1, 1971, a time when intercity service could be said to be on the verge of extinction. Rather than allow it to expire, Congress created a quasi-public enterprise and set as its goal the rejuvenation of rail travel on a profit-making basis. The Passenger Corporation is quasi-public because it is financed both privately, by the railroads that joined it, and publicly, through direct subsidy and guaranteed loans. The Secretary of Transportation chooses the endpoints to be served while the trains themselves are operated jointly by individual railroad and Amtrak employees.

Rail advocates praised the Passenger Service Act of 1970 which created Amtrak, while a rail executive probably summarized the industry's position when he said the legislation "is a sentimental excursion into the past for legislators over 50" that would not be at an excursion rate.[1] To date Amtrak has distinguished itself by losing over $100 million in each year of its existence, playing politics with train routings, running its trains late almost as much as on time, and providing on-train services that would be described flatteringly as inconsistent. However, the Corporation has accomplished something unexpected. It has reversed the decline in intercity rail travel. Fiscal 1973 revenue climbed 16 percent above the 1972 figure. Ridership in the 1973 calendar year is expected to show a second consecutive annual increase of 10 percent.[2]

I. Amtrak's Two Passenger Markets

Projecting a future for Amtrak entails predicting its chances for success in each of the two markets it serves. The first is the long-haul market for trips lasting a day or more over distances ranging from several hundred miles to coast-to-coast. The second is the corridors market, for relatively short trips lasting from one to five hours. If we assume a mile-per-minute average corridor speed, these involve distances of 60 to 300 miles.

Predictably, the requirements for public acceptance in the two markets differ. Amtrak's long-haul trains feature sleeping car accommodations, dining and lounge service, plus chaise-longue style coach seating. It is these services that Amtrak critics underscore as particularly pointless. Such services cannot compensate for the speed of air transport or the convenience of the automobile. Costwise, they are labor intensive and incapable of increases in productivity.[3] If properly priced, virtually all patronage of them eventually will be eliminated. The experience of the Canadian National supports this thesis. After it refurbished passenger services extensively in the early 1960's, its revenues doubled during 1962–1968.[4] But though a break-even operation was predicted, deficits never narrowed significantly.

---

* Associate professor of transportation, University of Tennessee.

[1] "That Lonesome Road."

[2] Albert R. Karr.

Despite these prospects, some of Amtrak's most popular and profitable routes are long distance runs. In June 1973 the profit on two of the longest routes, Chicago-Los Angeles and Chicago-Seattle, nearly equalled those earned on Northeast Corridor service, considered the Company's most lucrative operation.[5] The Seattle-Los Angeles route also showed a profit and the remaining transcontinental services almost broke even.

Significantly, Amtrak has raised fares on the service between Chicago and West Coast points by at least 10 percent since its inception, probably on the assumption that at least at certain periods of the year the demand for long distance train service is price inelastic. As for the future, the Department of Transportation, in a March 1973 report to Congress, predicts that several long-haul routes in fiscal year 1975 will be as profitable or lose as little as several of the corridors. New York-Florida and Chicago-Los Angeles operations are predicted to be the most profitable in the entire system.[6]

In view of the product offered by the Corporation, its long-haul performance is surprising. Equipment on the trains is of early postwar vintage and because of previous neglect has been prone to breakdowns resulting in delays and passenger discomfort. On-time performance has been poor; in August 1973, for example, long distance trains arrived on time (defined as within 6 minutes of the time shown in the timetable) only 18 percent of the time.[7] Although new equipment especially designed for passenger comfort is available, speed will at best remain the same in coming years as the railroads maintain their rights-of-way to meet freight rather than passenger train standards. For this reason, Amtrak has deemphasized speed in favor of the cruise-type appeal of train travel.

The Corporation at the same time has facilitated access to its trains through use of travel agents, car rentals, a nationwide reservation system and acceptance of credit cards—features long associated with air travel. Intermodal tour packages are promoted alongside all-rail offerings. Autoferry service is planned next year on three routes. The Company has enough confidence in its product, at least in some markets, to be introducing and/or scheduling fare increases of 10 percent and higher.[8]

Will this package be enough to lure passengers away from the speed of air travel or the convenience of the automobile in sufficient numbers to enable a profit-making long-haul rail service? Most observers are pessimistic. A popular argument is that only the old and/or those afraid to fly will ride the train. Yet surveys

---

[3] Gilbert Burck.

[4] Patton (1971).

[5] Karr.

[6] U.S. Department of Transportation (*DOT*) (1973).

[7] National Railroad Passenger Corp. (*NRPC*), Oct 18, 1973, p. 2.

[8] "Rail Travel Will Cost More."

of train ridership in the last two years by *Newsweek* and *U.S. News & World Report* reveal a growing attraction to rail travel by the younger generation.[9] A change in tastes for long distance travel by the coming middle-aged, middle-income segment of the population is a key to Amtrak's survival.

Nationally, Amtrak accounts for less than one percent of the intercity travel market. However, it holds six percent of this market where it does provide service.[10] On long hauls, is this sufficient to warrant continued subsidies? Conceptually, subsidy for rail is justified when the public expenditure to operate a rail service is less than the cost to build equivalent highway capacity for the service's displaced train riders. This argument does not apply to long distance trains since several times the number of passengers currently riding these trains could be absorbed by our airline and highway systems without increasing congestion.

In the long run, therefore, Amtrak's long haulers collectively should pay their own way. Furthermore, service should be self-supporting on a total cost basis, with Amtrak payments to contracting railroads reflecting full and not avoidable cost, the current basis of compensation under the Passenger Service Act. The major rail criticism of the Amtrak legislation is that the carriers actually lose money operating the trains. Thus, there is a disincentive to run them according to schedule when they delay the carriers' profitable freight services. The long-run period should include sufficient time to introduce technologically advanced locomotives and rolling stock to enable a more attractive product that at the same time can be operated at a lower cost than present equipment.

There are indications that the long haul market is turning around. If long-run profits are elusive despite high patronage levels, a difficult decision will be faced. Cutting off public funds will be politically unappetizing, yet lavishing perpetual subsidy upon the trains is not economically justified.

The corridors of the megalopolis present rail travel with its best opportunity to become a true alternative to automobile, air, and bus transport. For this to occur, trains must run at close to streetcar frequency, at speeds equal or faster than are feasible on parallel highways that, in turn, are subject to congestion all day. Here trains can compete in terms of transit time and cost and, most important, convenience. The relative inflexibility of fixed train departure times is offset by the frustration of highway congestion.

While long-haul rail service is argued to be too slow for business travel, corridor service must meet the requirements of this market. This brings it into competitive contact with both highway and air transportation. "British experience has shown that as trip time approaches three hours, rail begins to make substantial inroads on air traffic."[11] Presumably the three-hour time span represents an hour at both origin and destination getting to and from an airport, together with an hour's flying time. Whether there is an optimal competitive time relative to highway transport is not known; however, the thesis that variance in individual train revenues is a function of a three-hour station-to-station rail transit time is supported by personal research.[12] If three- to four-hour maximum end-points travel time plus frequent service can make rail competitive, one actual and four potential corridors can be distinguished.

Amtrak's single corridor is the 227-mile New York-Philadelphia-Washington route.

---

[9] "Railroads: The Youth Trip," and "People Crowd Trains Again, Find Problems—And Pleasures."

[10] "Viewpoint."

[11] *DOT* (1973), p. 52.

[12] Patton (1973).

North of Philadelphia 76 trains, excluding commuter runs, operate each weekday, 50 of them serving the entire route. Metroliners average 75 miles per hour between end-points while conventional trains approach a mile-a-minute average.[13] Thus, maximum end-points transit time is four hours. The potential corridors are: southern New England, eastern New York, and eastern Pennsylvania, with Chicago north to Milwaukee constituting the single possibility outside the Northeast. These routes have not attained corridor status because train frequency and/or speed does not yet produce an alternative to road transport in view of current levels of highway congestion.

Amtrak operates several additional relatively short-haul routes, but these do not fit the corridor mold. They resemble long-distance runs without the amenities. Conclusions regarding long-haul trains apply equally to these currently nonessential routes.

The true corridors, especially New York-Washington, deserve the opportunity to use effectively the advantages of the rail mode. Performance in this market has been very encouraging despite deficiencies connected with the service. For the price of two Boeing 747 jetliners, Penn Central initiated Metroliner service between New York and Washington in 1969.[14] Five years later average ridership for twelve weeks equals total first-year patronage. Since their introduction, these trains have carried 7.5 million passengers despite an on-time record of between 50 and 70 percent, limited food service and no lounge facilities, and a fare increase of 50 percent.[15] Conventional trains are equipped with some of the oldest cars in service. Collectively, however, the Metros and the conventionals account for 19 percent of the corridor's total travel market.[16]

Amtrak is purchasing fifty-seven Metroliner-type cars for this route. This will enable a 100 percent increase in Metro-type service and extend the concept to Boston.[17] It is a proper move but it is insufficient. To determine the effectiveness of rail travel in the New York-Washington corridor, the program outlined by the Department of Transportation in 1971 for passenger development in the Northeast should be implemented. After analyzing projected population growth and consequent travel demands in this area together with the capabilities of all available transport modes, the agency recommended a substantial upgrading of rail facilities. The program would eliminate remaining grade-crossings, fence the right-of-way, and build new track in rail-congested areas. Service would extend beyond New York and Washington into the suburbs and train frequency would increase.[18] The cost of track improvement would range from $100 to $500 million depending upon a goal of New York-Washington running time of two or two and a half hours. The report recommended $100 million be spent for new cars. This would buy 240 cars at current prices.[19]

A profit is predicted for the improved service; thus, initial federal assistance should be in the form of loans or guaranteed bonds as opposed to grants. Profits, in fact, would pay for the improvements that would enable a two-hour transit time. Though not quantified, a substantial increase in rail's share of this market is anticipated. Such a shift should eliminate the need for completely new highways between major cities and reduce the bill for improvements upon existing facilities. Current congestion at area airports should be alleviated as many short flights are curtailed and the carriers concentrate on long-haul services.

---

[13] *NRPC,* Oct. 28, 1973.

[14] One-fifth of the investment, the cost of the high-speed cars, was federally funded.

[15] *DOT* (1973); *NRPC,* Oct. 18, 1973, pp. 2–3, *NRPC,* Oct. 12, 1973, pp. 1–2.

[16] "Viewpoint," p. 54.

[17] *NRPC,* Oct. 12, 1973, p. 1.

[18] *DOT* (1971), *1,* 4–5, 22; *2,* Table 5A–3.

[19] C. D. Washington. Four hundred of the most comfortable cars used in Europe could be purchased for the same price.

Although profits are predicted, the question of the future of such a service if profits do not materialize must be confronted. The decision of whether or not to abandon the system should depend upon patronage and, if rail travelers are forced to use alternate transport modes, the public aid requirements of such alternatives. Public assistance to maintain a rail service is justified if it is less than that required to provide an adequate, equivalent service for displaced rail patrons. This argument justifies public support of commuter bus and train operations, and it should apply equally to heavily used intercity corridor services.[20] It constitutes the single exception to the rule that profit should be used to measure the long-run viability of the Corporation.

A final factor, the energy crisis, may present Amtrak with new market opportunities, presumably in both the long haul and the corridors. If fuel prices go higher, rail's attractiveness as a travel alternative should be enhanced, but this might be offset by increased use of car pools and the like; thus, it is impossible to forecast the net effect on patronage. Only the introduction of World War II-type rationing will produce a significant move to rail travel, and with Amtrak currently experiencing a car shortage, this shift would have to be accompanied by a crash program of equipment construction. Amtrak cannot base its long-run prospects on such imponderables.

II. Politics and Profit Constraints

In October 1973 President Nixon signed the Amtrak Improvement Act providing funding into 1974. In addition, however, the legislation contained three provisions that make Amtrak's search for solvency even more difficult than under the original 1970 ground rules. First, all existing routes are frozen for another year. Second, a new experimental route, to be selected by the Secretary of Transportation, is to be offered to the public each year. Third, Amtrak service standards will come under closer scrutiny by the Interstate Commerce Commission.[21]

Because Amtrak was created by and remains dependent upon Congress, it is perhaps inevitable that the legislators interfere in Corporate decision making. But rather than having to freeze routes and add new ones, the Company should be encouraged to abandon losing services and invest its limited resources in routes with the strongest profit potential. Greater flexibility, not less, is required in the Amtrak legislation. If Congress insists upon imposing money-losing (but politically popular) routes on the Corporation, the latter cannot be blamed if it does not meet the profit objective established for it.

To sum up: of the two markets served by Amtrak, the corridors represent its best chance to facilitate travel and earn a profit. Fast, frequent, relatively worry-free service provides a feasible alternative to highway and air transport. As in the case of commuter transit, long-run subsidy for corridor operations is justified if it is less than what the public must pay to provide alternative means for displaced rail patrons.

The remaining short hauls and the long distance trains eventually must become self-supporting. Amtrak gives the country its first chance to reap the benefits of a single, nationwide enterprise whose sole function is to provide rail passenger service. Patronage has been encouraging despite occasionally mediocre service. Notwithstanding the results of a widely heralded poll that "there is a clear and decisive mandate for providing and continuing and improving intercity passenger train travel," the strength of this market is unpredictable.[22]

---

[20] For an example see Government of Ontario Transit, pp. 11–12.

[21] Association of American Railroads; "The Amtrak Improvement Act of 1973."

Finally, if it is impossible to eliminate politics from the Corporation's decisions and such influence inhibits long-term profits, then as a society we must decide, in view of our heavy commitment to highway and air transport, how much we want to spend on a third travel alternative. Be it profits or politics or some combination of the two, we must determine ultimately the point of the pointless arrow.

REFERENCES

G. Burck, "Transportation's Troubled Abundance," *Fortune,* July 1971, *84,* 138.

A. R. Karr, "Amtrak's Loss Narrowed in Year Ended June 30," *Wall St. J.,* Oct. 15, 1973, 3.

E. P. Patton, "NRPC: Requiem or Renaissance?" *High Speed Ground Transportation J.,* Winter-Spring 1971, *5,* 88–91.

———, "AMTRAK—Some Interesting Correlations," *The Logistics and Transport. Rev.,* Fall 1973, *9,* 227–35.

C. D. Washington, "How to Self-Destruct," *Rail Travel Newsletter,* Oct. 1973, *3,* 7.

Association of American Railroads, "President Signs Amtrak Funding Bill," *Information Letter,* No. 2094, Nov. 7, 1973, 1–4.

Government of Ontario Transit, *GO Transit,* Toronto 1968.

National Railroad Passenger Corp., "Amtrak News," Oct. 12, 1973.

———, "Monthly On-time Performance Listing," Oct. 18, 1973.

———, *Amtrak National Schedules,* Oct. 28, 1973.

U.S. Department of Transportation, *Recommendations for Northeast Corridor Transportation,* Washington 1971.

———, *The Rail Passenger Service Act of 1970,* Washington 1973.

"Harris Poll Projects Big Future for Amtrak," *Railway Age,* Oct. 9, 1972, *173,* 39.

"People Crowd Trains Again, Find Problems —And Pleasures," *U.S. News & World Report,* July 2, 1973, *75,* 59–61.

"Rail Travel Will Cost More," *Rail Travel Newsletter,* Oct. 1973, *3,* 2–3.

"Railroads: The Youth Trip," *Newsweek,* Sept. 4, 1972, *80,* 61.

"That Lonesome Road," *Forbes,* Nov. 16, 1971, *108,* 34.

"The Amtrak Improvement Act of 1973," *Rail Travel Newsletter,* Oct. 1973, *3,* 3.

"Viewpoint," *Railway Age,* May 14, 1973, *174,* 54.

---

[22] "Harris Poll Projects Big Future for Amtrak."

The Wheel: Should We Reinvent It?

By HUGH S. NORTON*

The steel wheel was a major factor in *U.S.* intercity passenger transportation from roughly 1870 to 1950. The railroad's reign as premier common carrier was long and prosperous, but the decline was unmistakable after 1950.

Except for a handful of dedicated rail fans and elderly travelers, few found much fault with these developments. After all, is not the aircraft the epitome of mobility in a mobile nation? Yet on closer examination our intercity passenger transport system is less effective than it seems, or than we would like. While mobility between major cities has undoubtedly increased, that between major and secondary areas may have decreased in recent years. Like many of the complex problems in our modern society, this one is a result of a mixture of market forces, misguided public policies, and unforeseen circumstances, seemingly beyond solution.

Although the auto and the aircraft have, between them, come to dominate intercity transportation since 1950, we like to think that there is an alternative. We would like to think some fairly large portions of the intercity passenger market would be served more efficiently at less cost to the individual and to society by the railroad. However, as the mid–1970's approach, there seems to be very little justification for optimism.

To approach this problem, let us first look at the position of the railroad as a passenger carrier in recent decades. At the turn of the century the railroad was in a commanding position. As time passed, population and personal income grew and so did use of the automobile and airplane. But the rail share of common-carrier passenger miles, of course, declined. It rose markedly during the Second World War, but by 1950 it had fallen to 46.3 percent, declined to 27.8 percent in 1960, and by 1970 had sunk to about 7.5 percent.[1]

It must not be assumed that attempts to retain and reestablish rail passenger service have been entirely lacking. One may question the degree to which the railroads as an industry have been wholeheartedly behind these efforts, but they have, nonetheless, been made and at substantial cost. These efforts fall generally into two time periods.

The first was in the mid–1930's when the auto and the bus began to be recognized as real threats. This was the era of the "streamline" train when long-distance travel by rail was still reasonably competitive in terms of time (37 hours, Chicago to Los Angeles). Airlines were still primitive, and coast-to-coast travel required numerous changes and appealed mostly to those both venturesome and affluent. Although the railroads had by this time generally decided to abandon the short-haul traffic and were looking askance at the commuter train, the long haul was still attractive and appeared to be viable. The automobile in the 1930's was still a vehicle for local use. Though intercity highways were rapidly expanding, long-distance travel by auto was avoided by those who were able to afford Pullman travel. For comfort, dependability, and reasonable speed the first-class long-distance passenger train was a realistic competitor until the late 1930's.

---

* Johnson Professor of Economics, University of South Carolina.

[1] See Association of American Railroads 1971.

These efforts to hold or to gain passenger traffic were generally sincere. Major railroads did introduce streamline equipment with air-conditioning, low-cost food service, and other attractive features. To some degree these were effective, but technological advance in autos and aircraft, coupled with enormous outlays on highways and airports, rendered them largely impotent.

On the eve of the Second World War it was clear that the future of rail passenger service was in grave doubt. Rail travel won a temporary reprieve during the war due to limitations on gasoline, but as war ended the railroads, with cash reserves and diesel locomotives, began to look again at the future of passenger trains. Some, at least, were far from ready to abandon the effort.

The second serious attempt to resuscitate the passenger train was in the early 1950's. Again the western lines, in particular, made brave efforts.[2] Eastern railroads were much less involved although some, notably the New York Central, made valiant efforts with their name trains. Some routes throughout the nation seemed to retain vitality—the New York-to-Florida routes, for example, and a few transcontinental trains, such as the California Zephyr and the Santa Fe Super Chief, were popular money-making trains, but the list dwindled rapidly. By 1965 it had become almost impossible to travel substantial distances by trains since those which remained had few connecting trains still available.[3]

For the most part, these efforts came to naught. They were in some cases admittedly halfhearted, but in all cases they were carried out in the face of expanding interstate highway construction and ever increasing encouragement of aircraft and airport development. In the light of hindsight, they were doomed at the outset.

By the late 1960's the trend was clear beyond all doubt. A 1969 study by the Interstate Commerce Commission indicated that some eight railroads accounted for roughly 40 percent of the nation's noncommuter passenger business (Santa Fe, Chesapeake and Ohio-Baltimore and Ohio, Great Northern, Illinois Central, Missouri Pacific, Seaboard Coast Line, Southern, and Union Pacific). Avoidable costs for 1968 amounted to $118 million; that is, if carriers had not performed the business, they would have saved $118 million net. To put the matter another way, for every dollar in revenue *not earned* by *not performing* the business, they would have avoided the expenditure of $1.83! Further, these figures do not include sale or salvage value of equipment on property no longer required, an amount calculated to be $69.2 million. An index of the rapidity with which rail passenger service was losing ground was provided by a notation in the same report that in the one year, 1968–69, fifty trains had been withdrawn and fifty more were in the process of withdrawal. The report concluded:[4]

> Government subsidies or other forms of substantial Federal aid may well be necessary to continue the operation of intercity passenger trains. Should the public need for such service warrant retention of these trains that cannot be operated without significant losses, we would support a program of Federal aid to carriers. The first step, however, is to find out what service the public needs and how much that level of service would cost.

---

[2] It is interesting to note that the railroads also made efforts in the late 1940's and early 1950's to regain some lost branch-line traffic through the use of the rail diesel car. The RDC, a self-propelled unit, was attractive. Had it appeared in 1929 or 1939 instead of 1949, it might have made a real difference. Unfortunately it was at least a decade too late.

[3] The efforts of these years are well documented in the U.S. Senate hearings. See, also, Interstate Commerce Commission Report 1959.

[4] See U.S. Interstate Commerce Commission, July 1969.

Only a few bright spots were on the horizon, such as the Washington-New York Metroliners and the auto-train service from Washington to Florida, which appeared to have some likelihood of success. On the other hand, the whole railroad industry was shaken when the Penn Central collapsed in 1971.[5] A portent of the future was visible when the California Zephyr, one of the most popular of western trains, was withdrawn in early 1970.

I. What Does the Future Hold?

There are clearly large segments of the population under thirty years of age who have never been on a train (except, perhaps, on a short ride with Grandpa so they would be able to have the experience). Large numbers of those over forty are limited in their experience to military travel during the Second World War, which left them with a grim image of train travel and a strong determination never to repeat it. Certainly most of these people make their travel plans on a two-mode basis, i.e., automobile or plane. This is not to say that these decisions are always rational; clearly, they are often not, but they seem unlikely to change soon.

Despite this gloomy prognosis, one is reluctant to concede that the end of rail passenger service has finally arrived. By their nature railroads can offer much as passenger carriers in the modern world. Indeed, recent public policy developments relating to environmental protection, resource conservation, and energy use have made it more desirable to reconsider the merits of the railroad.

But the railroad can be resuscitated as a major factor in intercity passenger service only if it is or can be an attractive alternative. Public policy designed to force the use of rail travel or to coerce the passenger will not be successful. Let us turn our attention to the possibility of the passenger train's becoming, once again, a real competitor for intercity traffic.

II. The Railroad as a Competitor in the Passenger Field

Some passenger-carrying functions of the railroad clearly cannot compete and so, in most cases, can be ruled out immediately. Foremost of these is the trip in excess of 800 to 1,000 miles or more. Even with trains averaging over 100 miles per hour, such trips would take close to eight or ten hours. The airplane reduces this to three hours. If ground transportation time is taken into account, rail travel between heavily populated metropolitan areas over such distances might be feasible, but only, one suspects, in rare cases.

Another role of the railroad that can almost certainly be dispensed with is the short or intermediate distance trip (within the 300- to 400-mile range) between lightly populated areas. The private auto, the bus, and feeder airlines seem to be providing adequate service. The traffic congestion in and around airports and highways at these points is not serious. Even where the trip is from a lightly populated to a heavily populated area, the traffic congestion is at one end of the trip only. If railroads were able to increase their service, it would probably be at the expense of the airlines and the venture would likely be unprofitable for both, thus creating other public problems. Some of the major airlines have been hopeful that Amtrak will relieve them of some of their short unprofitable routes, but as yet small airlines seem to find them attractive and profitable.

Thus, it seems likely that the role of the railroad must be confined to trips roughly in the 500- to 1,000-mile range between heavily populated areas where traffic congestion is already severe and land use, smog, noise, etc., are serious factors. Only in these locations can the train lure enough passengers to be able to compete with other forms of transportation without continuous public aid.

---

[5] There were many aspects of the Penn Central problem which had nothing to do with passengers, but the overall impact was unfortunate.

When we consider the factors relating to the demand for the various modes of travel, for such distances it appears that the railroad might well have a decided advantage. The major factors are elapsed time, convenience (frequency of service, location of terminal, etc.), price, comfort (employee courtesy, meals, etc.), and dependability.

For intermediate distances, rail can compete and probably is equal or superior to either air or bus with respect to elapsed time when "downtown-to-downtown" movement is considered. This factor will be increasingly on the side of the railroads as airports move further away from the central city and become even more regional in nature and as highways become more crowded.

Convenience is, of course, a many-faceted concept. If most travelers would prefer to arrive and leave at the center of the city, railroads have the advantage with regard to terminal location.[6] Insofar as frequency of service is concerned, the railroad could presumably provide more frequent service than the airlines. This advantage of frequency also will increase as terminal air space becomes more critical. The price relationship for rail and air gives neither mode an advantage. Clearly, the rates for intermediate distances could be made comparable. Again, one must consider that costs of ground travel, both in money and time, to and from airports are increasing. Thus, for intermediate trips the ground transportation costs are a significant part of the total. On trips of up to 500 miles, the ground costs may be as much as 20 percent of the air fares and must be carefully considered as a part of the total trip cost. In almost every case ground costs for rail travel are much less both in money and in time. For some travelers costs of time are minor, but for business executives and professionals they may be very significant.

Comfort, employee courtesy, food service, etc., can also be comparable. Airlines once had a major advantage in this area, but they have lost much ground. The image of the railroads in this regard is, of course, very poor, but there is no reason why it might not improve at least to the low level conveniently set by the air carriers.[7]

The factor of dependability is clearly with the railroads. Mildly adverse weather, such as fog or light snow, which confounds air schedules and renders the auto almost useless, has little influence on rail operations. Aside from strikes, which might affect any common carrier mode, the advantage of dependable schedule operations should lie with the railroad.

In summary, there is no basic reason why rail service for intermediate distances between heavily populated areas cannot be made equal in almost all respects and superior in some to the airline and the bus. The private auto is another matter, but even here crowded freeways, costly parking, and other disadvantages are at work to undermine its long domination of the scene. Apparently the auto user will put up with fantastic amounts of hardship before he gives up, but there must be a limit somewhere. However, these competitive advantages and disadvantages are only part of the picture.

---

[6] Both those in their home city and those visiting probably would prefer to arrive downtown since the visitor generally wants to be there and the resident can make his way to the suburbs on his own. One cannot, of course, expect the carrier to maintain terminals in all areas of the metro. On the whole, the downtown location is best for all purposes.

[7] The "airbus" will make it even more difficult for the airlines to pay close attention to individual services, and problems of luggage handling, etc., will become even more intense inasmuch as most air terminals are already hard pressed, if not beyond capacity.

In the Northeast and in other highly urbanized areas of the nation, the impact of the auto and the aircraft in terms of space requirements, pollution, and use of energy resources has begun to raise serious objections to their further expansion. Ready solutions have not been obvious so far.

The comparative efficiency of the railroad as a mass carrier relative to resource use is beyond doubt. Perhaps we have reached a point where public policy considerations will outweigh private preferences in the market.

Are we, in fact, nearing a point where, after three decades of steady decline in the market for rail transport, passengers will be forced back to the rails because lack of space, pollution, or other nonpecuniary factors will have driven them unwillingly from their autos and airplanes? This seems to me both unlikely and undesirable. Even if these would-be riders presented themselves willingly or under duress at rail terminals, they could not be accommodated at the present time.

Rail passenger car inventory has declined alarmingly. Total *U.S.* passenger cars (including Pullman) have declined from 61,728 in 1929 to 6,842 in 1971. There were 9,469 Pullman cars in 1929 and less than 800 in 1971.[8] In recent years even this meager supply has been redundant.

Amtrak has acquired and upgraded many cars, but passenger-related physical facilities such as cars and stations have been allowed to decline rapidly. Since 1950 the railroads have abandoned their hopes for passenger service to the degree that any revival would require an almost complete reconstruction of the car fleet and many supporting facilities.

Any realistic revival of passenger service would not, of course, require the elaborate and costly long-distance equipment such as that of the "golden age" ending in 1930. Pullman, lounge, and elaborate dining cars would not be needed for runs of the distances likely to be competitive and compatible with public policy. Most of these runs would be in daylight hours, although some possibilities of attractive overnight service come to mind.[9]

On the other hand, many existing cars would have to be scrapped and replaced if an attractive image is to be created. Costs would be tremendous. It thus seems logical to argue that passenger service by rail, whatever its attractions, has a future only if the following developments occur.

1. Public and private costs of air and auto travel will continue to increase.
2. Air space, land use, pollution, fuel shortages, and other problems will continue to draw attention and perhaps force curtailment of auto and aircraft use and construction of facilities.
3. Funds are made available to upgrade stations, rights-of-way, etc., and to supplement the passenger car fleet with attractive equipment. Since railroads cannot undertake this effort, loans or grants will be essential.
4. A program of public education is undertaken aimed at the population group under fifty who have either a poor image or no image at all of rail travel.

These are large "if"s. It is not really very likely that all of these things will materialize in the near future.

Even under the best of circumstances it is difficult to revive an industry so far into decline as that of rail passenger service. A whole new generation must be made aware of alternatives. A half-century policy of promoting the highway and airway system must be slowed or reversed. Massive shifts in public taste and policy do not occur quickly.

---

[8] All data from Association of American Railroads 1972.

[9] For example, Los Angeles-San Francisco, possibly New York-Chicago, and some New York-to-Florida runs might attract such patronage. Most other runs would be completed in less than ten hours and no elaborate facilities would be needed.

Auto and airline interests will not abandon their present positions without considerable resistance, and their success in the legislative arena has been impressive in the past. Nor will these modes rest on their present technology. Improvements in efficiency, safety, and comfort will be introduced and exploited in the market.

Everyone agrees that *someone else* should drive less or fly less to ease the problems of resource use. Public concern in this matter has been given vocal expression for the past several years, but articulation of these views has far surpassed performance. Public criticism of expanded highway construction, for example, has been increasing, yet one notes with interest that auto registrations increased steadily during that period at a rate of about four million vehicles per year.

Likewise, lamentations are heard throughout the land about expanding airports, noise, and pollution, but air travel (revenue passenger miles) rose from 70.9 million in 1969 to 95.8 million in 1970. These figures hardly indicate a drastic reversal of public attitudes as regards the auto and the airline.

To summarize the situation: since 1930 a very comprehensive rail passenger system passed from the scene, largely as a result of government policies. But, after thirty years of intensive public aid, the wisdom of continued airline and highway development now begins to seem doubtful. Development of the rail alternative will require enormous restoration outlays and seems to have aroused very little public support. All in all, I am forced to take a grim view of the future of intercity rail passenger service.

REFERENCES

Association of American Railroads, *Yearbook of Railroad Facts*, Washington 1971.

———, *Yearbook of Railroad Facts*, Washington 1972.

Interstate Commerce Commission, *Report on Railroad Passenger Train Deficit*, Washington, May 1959, doc. no. 31954.

———, *Investigation of Costs of Intercity Rail Passenger Service*, Washington, July 1969.

U.S. Senate, Surface Subcommittee Transportation, *Hearings on Problems of the Railroads*, Washington 1957.

THE ECONOMICS OF THE FIRST AMENDMENT

The Market for Goods and the Market for Ideas

By R. H. COASE*

The normal treatment of governmental regulation of markets makes a sharp distinction between the ordinary market for goods and services and the activities covered by the First Amendment—speech, writing, and the exercise of religious beliefs—which I call, for brevity, "the market for ideas." The phrase, "the market for ideas," does not describe the boundaries of the area to which the First Amendment has been applied very exactly. Indeed, these boundaries do not seem to have been very clearly drawn. But there can be little doubt that the market for ideas, the expression of opinion in speech and writing and similar activities, is at the center of the activities protected by the First Amendment, and it is with these activities that discussion of the First Amendment has been largely concerned.

The arguments that I will be considering long antedate the passage of the First Amendment (which obviously incorporated views already held) and there is some danger for economists, although not necessarily for American lawyers, in confining our discussion to the First Amendment rather than considering the general problem of which it is a part. The danger is that our discussion will tend to concentrate on American court opinions, and particularly those of the Supreme Court, and that, as a result, we will be led to adopt the approach to the regulation of markets found congenial by the courts rather than one developed by economists, a procedure which already has gone a long way to ruin public utility economics and has done much harm to economic discussion of monopoly problems generally. This approach is confining in another way, since, by concentrating on issues within the context of the American Constitution, it is made more difficult to draw on the experience and thought of the rest of the world.

What is the general view that I will be examining? It is that, in the market for goods, government regulation is desirable whereas, in the market for ideas, government regulation is undesirable and should be strictly limited. In the market for goods, the government is commonly regarded as competent to regulate and properly motivated. Consumers lack the ability to make the appropriate choices. Producers often exercise monopolistic power and, in any case, without some form of government intervention, would not act in a way which promotes the public interest. In the market for ideas, the position is very different. The government, if it attempted to regulate, would be inefficient and its motives would, in general, be bad, so that, even if it were successful in achieving what it wanted to accomplish, the results would be undesirable. Consumers, on the other hand, if left free, exercise a fine discrimination in choosing between the alternative views placed before them, while producers, whether economically powerful or weak, who are found to be so unscrupulous in their behavior in other markets, can be trusted to act in the public interest, whether they publish or work for the *New York Times*, the *Chicago Tribune* or the Columbia Broadcasting System. Politicians, whose actions sometimes pain us, are in their utterances beyond reproach. It is an odd feature of this attitude that commercial advertising, which is often merely an expression of opinion and might, therefore, be thought to be protected by the First Amendment, is considered to be part of the market for goods. The result is that government action is regarded as desirable to regulate (or even suppress) the expression of an opinion in an advertisement which, if expressed in a book or article, would be completely beyond the reach of government regulation.

---

* University of Chicago.

This ambivalence toward the role of government in the market for goods and the market for ideas has not usually been attacked except by those on the extreme right or left, that is, by fascists or communists. The Western world, by and large, accepts the distinction and the policy recommendations that go with it. The peculiarity of the situation has not, however, gone unnoticed, and I would like to draw your attention to a powerful article by Aaron Director. Director quotes a very strong statement by Justice William O. Douglas in a Supreme Court opinion, a statement which is no doubt intended as an interpretation of the First Amendment, but which obviously embodies a point of view not dependent on constitutional considerations. Justice Douglas said: "free speech, free press, free exercise of religion are placed separate and apart; they are above and beyond the police power; they are not subject to regulation in the manner of factories, slums, apartment houses, production of oil and the like" (*Beauharnis* v. *Illinois*). Director remarks of the attachment to free speech that it is "the only area where *laissez-faire* is still respectable."

Why should this be so? In part, this may be due to the fact that belief in a free market in ideas does not have the same roots as belief in the value of free trade in goods. To quote Director again: "The free market as a desirable method of organizing the intellectual life of the community was urged long before it was advocated as a desirable method of organizing its economic life. The advantage of free exchange of ideas was recognized before that of the voluntary exchange of goods and services in competitive markets." In recent years, particularly, I think in America (that is, North America), this view of the peculiar status of the market for ideas has been nourished by a commitment to democracy as exemplified in the political institutions of the United States, for whose efficient working a market in ideas not subject to government regulation is considered essential. This opens a large subject on which I will avoid comment. Suffice it to say that, in practice, the results actually achieved by this particular political system suggest that there is a good deal of "market failure."

Because of the view that a free market in ideas is necessary to the maintenance of democratic institutions and, I believe, for other reasons also, intellectuals have shown a tendency to exalt the market for ideas and to depreciate the market for goods. Such an attitude seems to me unjustified. As Director said: "the bulk of mankind will for the foreseeable future have to devote a considerable fraction of their active lives to economic activity. For these people, freedom of choice as owners of resources in choosing within available and continually changing opportunities, areas of employment, investment, and consumption is fully as important as freedom of discussion and participation in government." I have no doubt that this is right. For most people in most countries (and perhaps in all countries), the provision of food, clothing, and shelter is a good deal more important than the provision of the "right ideas," even if it is assumed that we know what they are.

But leave aside the question of the relative importance of the two markets; the difference in view about the role of government in these two markets is really quite extraordinary and demands an explanation. It is not enough merely to say that the government should be excluded from a sphere of activity because it is vital to the functioning of our society. Even in markets which are mainly of concern to the lower orders, it would not seem desirable to reduce the efficiency with which they work. The paradox is that government intervention which is so harmful in the one sphere becomes beneficial in the other. The paradox is made even more striking when we note that at the present time it is usually those who press most strongly for an extension of government regulation in other markets who are most anxious for a vigorous enforcement of the First Amendment prohibitions on government regulation in the market for ideas.

What is the explanation for the paradox? Director's gentle nature does not allow him to do more than hint at it: "A superficial explanation for the preference for free speech among intellectuals runs in terms of vertical interests. Everyone tends to magnify the importance of his own occupation and to minimize that of his neighbor. Intellectuals are engaged in the pursuit of truth, while others are merely engaged in earning a livelihood. One follows a profession, usually a learned one, while the other follows a trade or a business." I would put the point more bluntly. The market for ideas is the market in which the intellectual conducts his trade. The explanation of the paradox is self-interest and self-esteem. Self-esteem leads the intellectuals to magnify the importance of their own market. That others should be regulated seems natural, particularly as many of the intellectuals see themselves as doing the regulating. But self-interest combines with self-esteem to ensure that, while others are regulated, regulation should not apply to them. And so it is possible to live with these contradictory views about the role of government in these two markets. It is the conclusion that matters. It may not be a nice explanation, but I can think of no other for this strange situation.

That this is the main explanation for the dominance of the view that the market for ideas is sacrosanct is certainly supported if we examine the actions of the press. The press is, of course, the most stalwart defender of the doctrine of freedom of the press, an act of public service to the performance of which it has been led, as it were, by an invisible hand. If we examine the actions and views of the press, they are consistent in only one respect: they are always consistent with the self-interest of the press. Consider their argument that the press should not be forced to reveal the sources of its published material. This is termed a defense of the public's right to know—which is interpreted to mean that the public has no right to know the source of material published by the press. To desire to know the source of a story is not idle curiosity. It is difficult to know how much credence to give to information or to check on its accuracy if one is ignorant of the source. The academic tradition, in which one discloses to the greatest extent possible the sources on which one relies and thus exposes them to the scrutiny of one's colleagues, seems to me to be sound and an essential element in the search for truth. Of course, the counterargument of the press is not without validity. It is argued that some people would not express their opinions honestly if it became known that they really held these opinions. But this argument applies equally to all expressions of views, whether in government, business, or private life, where confidentiality is necessary for frankness. However, this consideration has commonly not deterred the press from revealing such confidences when it was in their interest to do so. Of course, it would also impede the flow of information to reveal the sources of the material published in cases in which the transmission of the information involved a breach of trust or even the stealing of documents. To accept material in such circumstances is not consistent with the high moral standards and scrupulous observance of the law which the press expects of others. It is hard for me to believe that the main thing wrong with the Watergate affair was that it was not organized by the *New York Times*. I would not wish to argue that there are not conflicting considerations in all these cases which are difficult to evaluate. My point is that the press does not find them difficult to evaluate.

Consider another example which is in many ways more striking: the attitude of the press to government regulation of broadcasting. Broadcasting is an important source of news and information; it comes within the purview of the First Amendment. Yet the program content of a broadcasting station is subject to government regulation. One might have thought that the press, devoted to the strict enforcement of the First Amendment, would have been constantly attacking this abridgment of the right of free speech and expression. But, in fact, they have not. In the forty-five years which have passed since the formation of the Federal Radio Commission (now transformed into the Federal Communications Commission), very few doubts about the policy have been expressed in the press. The press, which is so anxious to remain unshackled by government regulation, has never exerted itself to secure a similar freedom for the broadcasting industry.

Lest you think that I manifest a hostility to the American press, I would like to point out that the British press has acted in a similar fashion. In this case the contrast between actions and proclaimed beliefs is even stronger since what was established in Britain was a government-controlled monopoly of a source of news and information. It might have been thought that this affront to the doctrine of freedom of the press would have appalled the British press. It did not. They supported the broadcasting monopoly, mainly, as far as I can see, because they saw the alternative to the British Broadcasting Corporation (*BBC*) as commercial broadcasting and, therefore, as involving increased competition for advertising revenue. But if the press did not want competition for advertising revenue, they also did not want increased competition in the supply of news. And so they did their best to throttle the *BBC*, at least as a purveyor of news and information. When the monopoly was originally established (when it was still the British Broadcasting *Company*), the *BBC* was prohibited from broadcasting news and information unless obtained from certain named news agencies. No news could be broadcast before 7 p.m. and broadcasts likely to affect adversely the sale of newspapers faced other restrictions as well. Gradually, over the years, these restrictions were relaxed as a result of negotiations between the press and the *BBC*. But it was not until after the outbreak of World War II that the *BBC* broadcast a regular news bulletin before 6 p.m.[1]

But, it may be argued, the fact that businessmen are mainly influenced by pecuniary considerations is no great discovery. What else would one expect from the money-grubbers of the newspaper world? Furthermore, it may be objected, because a doctrine is propagated by those who benefit from it does not mean that the doctrine is unsound. After all, have not free speech and a free press also been advocated by high-minded scholars whose beliefs are determined by what is true rather than by more sordid considerations? There has surely never been a more high-minded scholar than John Milton. As his *Areopagitica* "for the liberty of unlicensed printing" is probably the most celebrated defense of the doctrine of freedom of the press ever written, it seemed to me that it would be worthwhile to examine the nature of his argument for a free press. Milton's work has another advantage for my purpose. Written in 1644, that is, long before 1776, we can see the character of the argument before there was any general understanding of how competitive markets worked and before the emergence of modern views on democracy.

---

[1] For a discussion of the attitude of the press to the monopoly of British broadcasting, see Coase, pp. 103–10 and 192–93.

It would be idle for me to pretend that I could act as a guide to Milton's thought. I know too little of seventeenth century England and there is much in Milton's pamphlet the meaning of which I cannot discern. Yet, there are passages which leap across the centuries and for whose interpretation no scholarship is needed.

As one would expect, Milton asserts the primacy of the market for ideas: "Give me the liberty to know, to utter, and to argue freely according to conscience, above all liberties" (p. 44). It is different from the market for goods and should not be treated in the same way: "Truth and understanding are not such wares as to be monopolised and traded in by tickets and statutes and standards. We must not think to make a staple commodity of all the knowledge in the land, to mark and license it like our broadcloth and our woolpacks" (p. 29). The licensing of printed material is an affront to learned men and to learning: "When a man writes to the world, he summons up all his reason and deliberation to assist him; he searches, mediates, is industrious, and likely consults and confers with his judicious friends; after all which done he takes himself to be informed in what he writes, as well as any that writ before him. If in this the most consummate act of his fidelity and ripeness no years, no industry, no former proof of his abilities can bring him to that state of maturity as not to be still mistrusted and suspected, unless he carry his considerate diligence, all his midnight watchings . . . to the hasty view of an unleisured licenser, perhaps much his younger, perhaps far his inferior in judgment, perhaps one who never knew the labour of book-writing, and, if he be not repulsed or slighted, must appear in print like a puny with his guardian and his censor's hand on the back of his title to be his bail and surety, that he is no idiot or seducer, it cannot be but a dishonour and derogation to the author, to the book, to the privilege and dignity of learning" (p. 27). Licensing is also an affront to the common people: "Nor is it to the common people less than a reproach; for if we be so jealous over them, as that we dare not trust them with an English pamphlet, what do we but censure them for a giddy, vicious, and ungrounded people, in such a sick and weak state of faith and discretion, as to be able to take nothing down but through the pipe of a licenser" (p. 30). In the market for ideas, the right choices are made: "Let [truth] and falsehood grapple; who ever knew Truth put to the worse in a free and open encounter" (p. 45). Those who undertake the job of licensing will be incompetent. A licenser should be, according to Milton, "studious, learned, and judicious." But this is not what we are likely to get: "we may easily foresee what kind of licensers we are to expect hereafter: either ignorant, imperious, and remiss, or basely pecuniary" (p. 25). The licensers are more likely to suppress truth than falsehood: "if it come to prohibiting, there is aught more likely to be prohibited than truth itself; whose first appearance to our eyes bleared and dimmed with prejudice and custom is more unsightly and unplausible than many errors . . . " (p. 47). Nor does Milton fail to tell us that the licensing scheme against which he was writing came about as a result of industry pressure: "And how it got the upper hand . . . there was in it the fraud of some old patentees and monopolisers in the trade of bookselling" (p. 50).

In the formation of Milton's views, self-interest may perhaps have played a part, but there can be little doubt that his argument embodies a good deal of intellectual pride of the kind to which Director refers. The writer is a learned man, diligent and trustworthy. The licenser would be ignorant, incompetent, and basely motivated, perhaps "younger" and "inferior in judgment." The common man always chooses truth as against falsehood. The picture is a little too one-sided to be wholly convincing. And if it has been convincing to the intellectual community (and apparently it often has), it is surely because people are easily persuaded that what is good for them is good for the country.

I do not believe that this distinction between the market for goods and the market for ideas is valid. There is no fundamental difference between these two markets and, in deciding on public policy with regard to them, we need to take into account the same considerations. In all markets, producers have some reasons for being honest and some for being dishonest; consumers have some information but are not fully informed or even able to digest the information they have; regulators commonly wish to do a good job, and though often incompetent and subject to the influence of special interests, they act like this because, like all of us, they are human beings whose strongest motives are not the highest.

When I say that the same considerations should be taken into account, I do not mean that public policy should be the same in all markets. The special characteristics of each market lead to the same factors having different weights, and the appropriate social arrangements will vary accordingly. It may not be sensible to have the same legal arrangements governing the supply of soap, housing, automobiles, oil, and books. My argument is that we should use the same *approach* for all markets when deciding on public policy. In fact, if we do this and use for the market for ideas the same approach which has commended itself to economists for the market for goods, it is apparent that the case for government intervention in the market for ideas is much stronger than it is, in general, in the market for goods. For example, economists usually call for government intervention, which may include direct government regulation, when the market does not operate properly—when, that is, there exist what are commonly referred to as neighborhood or spillover effects, or, to use that unfortunate word, "externalities." If we try to imagine the property rights system that would be required and the transactions that would have to be carried out to assure that anyone who propagated an idea or a proposal for reform received the value of the good it produced or had to pay compensation for the harm that resulted, it is easy to see that in practice there is likely to be a good deal of "market failure." Situations of this kind usually lead economists to call for extensive government intervention.

Or consider the question of consumer ignorance which is commonly thought to be a justification for government intervention. It is hard to believe that the general public is in a better position to evaluate competing views on economic and social policy than to choose between different kinds of food. Yet there is support for regulation in the one case but not in the other. Or consider the question of preventing fraud, for which government intervention is commonly advocated. It would be difficult to deny that newspaper articles and the speeches of politicians contain a large number of false and misleading statements —indeed, sometimes they seem to consist of little else. Government action to control false and misleading advertising is considered highly desirable. Yet a proposal to set up a Federal Press Commission or a Federal Political Commission modeled on the Federal Trade Commission would be dismissed out of hand.

The strong support enjoyed by the First Amendment should not hide from us that there is, in fact, a good deal of government intervention in the market for ideas. I have mentioned broadcasting. But there is also the case of education, which, although it plays a crucial role in the market for ideas, is subject to considerable regulation. One might have thought that those who were so anxious to obstruct government regulation of books and other printed material would also find such regulation in the field of education obnoxious. But, of course, there is a difference. Government regulation of education commonly accompanies government financing and other measures (such as compulsory school attendance) which increase the demand for the services of intellectuals and, therefore, their incomes. (See E. G. West, p. 101.) So self-interest, which, in general, would lead to support for a free market in ideas, suggests a different attitude in education.

Nor do I doubt that detailed study would reveal other cases in which groups of practitioners in the market for ideas have supported government regulation and the restriction of competition when it would increase their incomes, just as we find similar behavior in the market for goods. But interest in monopolizing is likely to be less in the market for ideas. A general policy of regulation, by restricting the market, would have the effect of reducing the demand for the services of intellectuals. But more important, perhaps, is that the public is commonly more interested in the struggle between truth and falsehood than it is in the truth itself. Demand for the services of the writer and speechmaker depends, to a considerable extent, on the existence of controversy—and for controversy to exist, it is necessary that truth should not stand triumphant and alone.

Whatever one may think of the motives which have led to the general acceptance of the present position, there remains the question of which policies would be, in fact, the most appropriate. This requires us to come to some conclusion about how the government will perform whatever functions are assigned to it. I do not believe that we will be able to form a judgment in which we can have any confidence unless we abandon the present ambivalence about the performance of government in the two markets and adopt a more consistent view. We have to decide whether the government is as incompetent as is generally assumed in the market for ideas, in which case we would want to decrease government intervention in the market for goods, or whether it is as efficient as it is generally assumed to be in the market for goods, in which case we would want to increase government regulation in the market for ideas. Of course, one could adopt an intermediate position—a government neither as incompetent and base as assumed in the one market nor as efficient and virtuous as assumed in the other. In this case, we ought to reduce the amount of government regulation in the market for goods and might want to increase government intervention in the market for ideas.

I look forward to learning which of these alternative views will be espoused by my colleagues in the economics profession.

REFERENCES

R. H. Coase, *British Broadcasting, A Study in Monopoly*, Cambridge, Mass. 1950.

A. Director, "The Parity of the Economic Market Place," *J. of Law and Econ.*, Oct. 1964.

J. Milton, *Areopagitica, A Speech for the Liberty of Unlicensed Printing*, with introduction and notes by H. B. Cotterill, New York 1959.

E. G. West, "The Political Economy of American Public School Legislation," *J. of Law and Econ.*, Oct. 1967.

Beauharnis v. *Illinois*, 343 U.S. 250, 286, 1952.

Regulation and Diversity in Communications Media

By William F. Baxter*

Our prevailing attitude toward freedom of the press—that the press should be totally independent of government regulation or direct influence—evolved during a time when the number of cities which enjoyed competing journalistic voices was many times as great as is the case today.[1] No significant change in that libertarian attitude is now evident, notwithstanding this decline in newspaper voices. No authority is recognized on the part of either the state or federal government to assure diversity, fairness, or balance in the presentation of news or editorial opinion by the printed media; and, with only trivial exceptions described hereafter, no obligation is recognized on the part of the printed media to make their pages available, even at a fee, to those who would like to express contrary views.

The history of the electronic media is quite different. The propagation of radio communication has been licensed by the Federal government since 1912.[2] Inexpensive receiving sets became generally available in the early 1920's, and private use of the electromagnetic spectrum for personal and commercial communication grew enormously in those years. Efforts by the Secretary of Commerce, acting pursuant to the 1912 legislation, to enforce restrictive licenses and to encourage self-regulation by industry with a view to reducing signal interference were largely unsuccessful. Congress responded with the Radio Act of 1927 (Act of Feb. 23, 1927, *STAT*, *44*, ch. 169, p. 1162).

The problem confronted was not significantly different from the most classic of all legal problems, namely, the problem of describing the multitude of interests which have been recognized with respect to the surface of the earth. One person may own mineral rights; another, as a typical lessee, may have the right for a period of years to use the surface itself and the immediately adjoining air space; still another may own the permanent or "fee" interest in the property itself. Subject to minimal restrictions, each of these interests may be privately transferred by the party who holds it; and each party is expected to—and generally does—enforce such ownership rights as he has through the courts of law at his own expense. A similar system of private ownership and private enforcement might have been adopted with regard to the radio spectrum, the initial allocation being made on the basis of historical priority as to use, unused portions being left subject to future private appropriation, as was done with the unclaimed lands of the Western territories (see, e.g., A. DeVany et al.).

In 1927, however, Congress took a very different approach. It created a Federal Radio Commission (*FRC*), made a license from the Commission a prerequisite to use of any portion of the spectrum, and made it clear that a licensee did not receive ownership of a described portion of the spectrum over a described geographic area, but, rather, obtained a mere right to use that portion of the spectrum, ownership remaining "in the public." Thus, over the entire history of broadcasting, an agency of the federal government has dispensed the right to use the electromagnetic spectrum in accordance with what it perceived, from time to time, to be "the public interest." That deplorably vague criterion constitutes all the guidance the statutes give.[3]

---

* Professor of law, Stanford University.

[1] Since 1910 the percentage of American cities with two or more newspapers has fallen from about 60 percent to less than 4 percent and the percentage of total daily circulation produced by newspaper chains has increased from about 10 percent to about 60 percent (E. Emery). No sinister plot is implied. There is persuasive evidence that scale economies in the production and distribution of newspapers have been the major factor in bringing about this sharp decline in the number of competing journalistic voices (J. Rosse).

[2] See Act of Aug. 13, 1912, Statutes at Large (*STAT*), *37*, ch. 287, p. 302.

The administering agencies, first the *FRC* and then its successor, the Federal Communications Commission (*FCC*), by their opinions in licensing cases and by their regulations, have given the criterion such content as it now has. And in the context of journalistic balance, the implications of the criterion are administered under the rubric of the "fairness doctrine."

The *FRC* first invoked the "fairness doctrine" in 1929 when it rejected a prior user's application in favor of a new applicant because the former had broadcast only one point of view. Asserting that the spectrum should not be used to allow expression of one "school of thought, religious, political, social, and economic," to the exclusion of others, it ruled that the continuous expression of a single point of view was not consistent with "the public interest."[4] This theme was continued and obscured by the *FCC* through the 1930's and early 1940's with the result that broadcasters totally abandoned the practice of editorializing and dropped much programming that might have been thought controversial. Unhappy with this development, the *FCC* conducted an inquiry during 1947–49 and finally ordained in a 1949 report that every licensee must devote a reasonable portion of broadcast time to the discussion of controversial issues and had a responsibility for the presentation of all sides of such issues (*FCC*, *13*, 1246). Because of the vagueness of this requirement and because the "death penalty" of refusing to renew the broadcaster's license was the only sanction which the *FCC* was able to invoke if it perceived that these responsibilities had not been discharged, industry concern was substantial. Despite a strenuous but not very successful effort at clarification by the *FCC* in 1964 (the "Fairness Primer," Pike and Fisher Radio Regulation, 2d ser., *2*, p. 1901), the responsibility of broadcasters remains obscure.

The situation was improved—or made worse, depending on one's point of view—when Congress in 1960 granted the *FCC* the power to impose fines on a broadcaster who "willfully or repeatedly fails to observe" any rule of the Commission. Precisely because this was a far less extreme sanction than that of license termination, this amendment created a possibility of far more frequent intervention by the *FCC*. In 1967 the *FCC* promulgated a more specific set of rules creating the right to reply to (1) *ad hominem* attacks on an identified person or group and to (2) licensee endorsements of or opposition to a legally qualified candidate for any office (*Code of Federal Regulation*, 1967, *47*, §73.123).

In this context, one should take note of the substantively related rule regarding "equal time." This requires that a licensee, if he covers any one declared candidate for office, afford equal time on equal terms to each other candidate for the same office (*USC*, 1934, *47*, §315a). Unlike the "fairness doctrine," this rule rests on an explicit provision in the Communications Act. But one should not infer from this difference in statutory basis that the

---

[3] See United States Code (*USC*), *47*, §§307a, 309a,

[4] See the Great Lakes Broadcasting decision, cited in the Red Lion case in Supreme Court Reporter, *89*, 1799.

"fairness doctrine" is beyond the *FCC*'s legal authority; in various ways Congress has taken note of and ratified the Commission's "fairness" requirements (*STAT*, 1959, *73*, 557). Rather, the point is that, apart from the *FCC*'s regulations regarding personal attack and implementing the statutory "equal time" rule, the more sweeping implications of the "fairness doctrine" cannot be found in any single written document but must be inferred from a series of rather obscure opinions.

The *FCC*'s attitude toward enforcement of the "fairness doctrine" has been highly permissive, and the number of instances in which it has attempted to enforce any sanction have been few. The most important of those instances was the Red Lion Broadcasting case. The Red Lion Company had broadcast over its Pennsylvania radio station a statement by the Rev. Billy James Hargis, as part of his "Christian Crusade," criticizing Fred J. Cook and his book, *Goldwater—Extremist on the Right*, which was critical of U.S. Senator Barry Goldwater. Hargis' statement, greatly condensed and somewhat oversimplified, was to the effect that Cook was a self-confessed liar and a Communist sympathizer. Cook demanded of the station that he be provided free time in which to reply. The station refused and Cook carried his complaint to the *FCC*, which ordered the station to provide him with free time, an order thereafter appealed to and upheld by the lower Federal courts.

While the Red Lion case was making its way through the lower Federal courts, the *FCC* promulgated regulations particularizing the obligation of broadcasting stations to afford reply time in instances involving personal attack (Federal Register, 1968, *33*, 5362). An association of news broadcasters challenged the constitutionality of those regulations in a different lower Federal court, where they were held unconstitutional. The Supreme Court then agreed to review simultaneously these two inconsistent lower court decisions and upheld both *FCC* actions, affirming one court's decision and reversing the other. The Court justified this rather extraordinary governmental intervention in the information-dissemination process by noting in its decision that substantially more individuals would like to broadcast their views than there are frequencies available to allocate (United States Reports, 1969, *395*, 367). It continued:

> Because of the scarcity of radio frequencies, the government is permitted to put restraints on licensees in favor of others whose views should be expressed on this unique medium. But the people as a whole retain their interest in free speech by radio and their collective right to have the medium functioning consistently with the end purposes of the First Amendment. . . . It is the purpose of the First Amendment to preserve an uninhibited market place of ideas . . . rather than to countenance monopolization of that market, whether it be by the government itself or a private licensee. [p. 390]

It is not obvious from a review of this history how best to describe the government objective to which the tradition of free speech is being subordinated. I believe that one puts the philosophically turgid pronouncements of the *FCC* and the Court in their best light by identifying the objective as balance and moderation on the part of *each* broadcaster. This seems to be implied by the obvious fact that quite different policies, involving a lesser departure from free speech tradition, would be appropriate if one were seeking station-by-station diversity which yielded an aggregate pattern of balance. Balance on *each* station is required, the Court seems to be saying, because spectrum scarcity precludes a large enough number of diverse voices to yield aggregate balance.

Not so, respond the critics of the Court: The spectrum is not so limited, at least in comparison with the newspaper field where First Amendment traditions still prevail.

What are the merits of these conflicting positions? First, the Court is on weak foundations in invoking spectrum scarcity as the primary justification. Even if one accepts the proposition that the electromagnetic spectrum appropriate for broadcast uses is finite and, hence, that the supply would become infinitely inelastic at some sufficiently high price, supply is not wholly inelastic at the prevailing price. There is no a priori reason to think that the supply elasticity of the electronic media is either greater or less than that of the aggregate of inputs prerequisite to the distribution of printed current event messages.

Indeed, it is not apparent that supply elasticities at any particular price range are relevant. The number of message-originating sources to which the typical citizen has access, as a consequence of the present demand-supply equilibrium, is a more relevant consideration, and that number has only a tenuous connection to spectrum "scarcity."

I. Media Regulation: A Paradox?

The obvious fact, then, is that the typical citizen is served by a substantially larger number of electronic media voices than of newspaper voices. Most cities have three or more television stations; in the top fifty markets the average number is about five with at least one of those being a UHF station; and, in addition, the typical city is served by dozens of AM and FM radio stations. Moreover, it is fairly clear that if the *FCC*'s fanatical attachment to the doctrine of localism were to be abandoned, the number of VHF stations available to the typical city could be increased by at least one and perhaps two, for more powerful, regional stations would not require that so much spectrum be "wasted" to avoid interference. In contrast, the typical city has available to it only one daily newspaper.

Nevertheless, in the context of the printed media, including the daily newspaper, our historic First Amendment attitudes prohibiting all governmental intervention prevail. But for electronic media, where the problem of concentration, or fewness of voices, is less acute, government intervention—justified on grounds of "scarcity"—is now established.

This situation has been called a deplorable paradox. Certainly it does have elements of paradox. And I do not wish to be understood, in what follows, to be asserting that the paradox is merely apparent or, less still, that readily available policies would serve to increase the diversity of journalistic voices. Nevertheless, it seems to me that the element of paradox is easily overdrawn.

The first serious error is made, it seems to me, when it is assumed that the appropriate data for comparison are the number of broadcasters on the one hand and the number of newspapers on the other. This assumption disregards critical differences between these media. In a twenty-four-hour day, one broadcaster can transmit about 200,000 words in normal patterns of speech; I estimate that, depending on length, a newspaper is the equivalent of anything between one to three broadcasters, with the Sunday *New York Times* equivalent to God knows how many.

But even this verbal capacity measure seriously understates the comparative advantage of the newspaper, for it has two other characteristics that the broadcaster lacks: random access and susceptibility to scanning.

By the former, I mean, for example, that one may read a newspaper at will, looking at the financial pages of the newspaper before dinner, the rest of the news after dinner, and skipping the sports, food, etc., pages entirely. But one must watch a 6 p.m. news program at 6 p.m. from the beginning.

Random access, though, is not the whole of the matter: for without susceptibility to scanning, the value of random access is greatly reduced. To illustrate, again by example, I may judge from a newspaper headline that the subject of the story may be of great or little interest to me, depending on the level of detail and sophistication with which it is written. I can sample the first and fifth paragraphs to get some notion of this in a matter of seconds and can then decide whether to read the item at any of several speeds or not to read it at all. In short, each individual part of the several "twenty-four-hour broadcasts" in a daily newspaper is practically available during prime (or any other) time. Analogous approaches to electronic media are not possible.

These media differences are important for the question of diversity precisely because these advantageous characteristics make it possible for the newspaper publisher to offer something for everyone. And this important potential of the daily newspaper has not gone unused: the typical large city daily has a marked propensity to publish a wild diversity of materials, often grouped in physically separate sections of the paper. In truth, it is not one newspaper, but a large number of alternative newspapers, any one of which an individual reader can compose for himself and tailor to his own interests and time constraints. Thus, the capacity of a single newspaper to deliver diversity, including diversity along the political spectrum through an assortment of syndicated columnists, greatly exceeds any analogous capacity on the part of a broadcaster.

And the diversity of the wide range of printed media—from specialist journals and periodicals to general-interest news weeklies—blurs the sharpness of the asserted paradox still further.

On the other hand, developments such as cable television may augment diversity among the electronic media. Although cable technology is largely at hand, I think it right to refer only to *prospects* of diversity—perhaps to remote prospects—rather than actualities. For example, it is wholly unrealistic politically to suppose that advertiser-supported, i.e., "free," over-the-air television will be permitted to disappear or even to lose through competition any substantial part of its present program material.

Given that fact, the costs of substantial penetration by cable will not be borne by subscribers until cable can confer utilities greatly in excess of those presently conferred by "free" TV. Sharper signals will not be enough. It is doubtful that the opportunity to import distant signals would be enough; and even that will not occur until some substantial change in the regulatory environment leads the *FCC* to abandon its present commitments to the concept of localism and to the notion that the capitalized rents of local broadcast licensees must be protected. Neither of these regulatory developments appears to be on the horizon. In my judgment, based on the studies to date, particularly the recent book by Roger Noll, Merton J. Peck, and John McGowan, nothing short of national networking among cable systems will do the trick. For the substantial costs of program origination must be spread over very large numbers of viewers; and large numbers with high utilities for programs of types not presently found on "free" TV cannot be assembled on anything narrower than a regional or perhaps a national basis. On the other hand, extensive networking among cable systems would pose such a serious threat to the survival of "free" TV that I do not expect it to be permitted soon.

Hence, I expect that we are stuck for a significant period of time with about the same diversity and number of editorial voices available to us today. Because I see no significant differences between the electronic and the printed media with respect to their capacity for diversity (the greater number of electronic originators being compensated for by the printed media's susceptibility to scanning and random access), I see no justification for significantly different rules regarding government intervention for the purpose of achieving diversity. Shall we, then, have government intervention in both categories of media or in none?

II. What Kind of Intervention?

Needless to say, I would opt for intervention if I could only be sure that I, or someone with my wisdom and insight, would be exerting the influence. But such awesome power lodged in other hands would be unsafe; and even if the power were not abused in the ordinary sense of that word, it is likely to be trivialized as I think is the case with the "fairness doctrine" as administered by the *FCC*. If the *FCC*'s frequent references to the doctrine in the abstract are taken seriously, one must conclude it is a pervasive and demanding requirement that all sides of an issue be presented if the issue is treated at all. Yet the abstract doctrine has so rarely been invoked with operational consequences (either fines or license denials) that one has difficulty taking the abstraction seriously. Some broadcasters appear to ignore it; others accord it ritualistic observance by broadcasting "opposing views" selected on the basis of their absurdity. As the court and administrative rulings discussed earlier illustrate, the occasional operational invocation of the doctrine has promoted not diversity but homogeneity.

More vigorous, purposeful intervention is a frightening prospect unless the levers are in hands of saints with great wisdom, and such men are in very short supply, particularly in government agencies. Deplorable as both sets of media are and have been in my judgment, I see no basis to hope that government intervention would improve that performance even slightly. To the extent there is a problem with the media, it is the same with respect to both sets: virtually total domination by a middle-of-the-road, middle-class point of view, articulated at a cocktail party level of sophistication, and slicked over with the superficial moralisms of journalism schools Why this should be so is not hard to ascertain: it represents that which will offend the smallest number among the audience. It is a manifestation of the now-familiar problem of audience division explained twenty-odd years ago by Peter Steiner. That phenomenon is exhibited by the mass-circulation printed media, as well as by the electronic media, for although subscription fees are charged, these fees are dominated as a revenue source by advertising fees which are a function of circulation. It is at least consistent with this causal ascription to suggest that the problem seems to be at its worst on network television. For it is this segment of the media which is most severely confronted with the need to appeal to a mass audience whose members are able to register only marginal preferences through an on/off switch and have no market in which to register the intensity of their preferences.

The spokesmen for that particular subsection of the media have even converted the difficulty to a virtue in their own minds: Eric Sevareid, a newscaster for the Columbia Broadcasting System, has warned ominously of the danger of

> moving toward the condition of France before World War II when I lived there. It was a nation of critics. Government was conducted in the front pages of the

> papers, about 47 of them, published in Paris every day. Nothing could get done. It was a kind of political/intellectual anarchy. So I would say that the immediate danger . . . is not the possible restrictions of communications monopoly but the reverse—the degeneration of public affairs journalism into a kind of Tower of Babel. [pp. 54–61]

There is every reason to think that government intervention would not lessen but rather exacerbate this problem, for the dominant pressure of a mass democracy has very much the same features as audience division on a free medium: a struggle to capture the middle, waged at a level of sophistication comprehensible to the middle, in which minorities are easily ignored. Thus government intervention, even under the aegis of a benign administration, in that part of the media which attempts to achieve a broad appeal constitutes carrying coal to Newcastle. As applied to more specialized segments of the media, intervention has in the past and probably will continue to suppress diversity. As to both segments, it carries a threat of worse consequences if administration should cease to be benign.

If, however, the historic attitude toward a free press is not to spread from the printed to the electronic media, if government intervention in both media is the wave of the future, then I would be inclined to argue that the second-best situation might well be represented by a "fairness" type of intervention in electronic media and an "access" type intervention for the printed media. The arguments against an access-type intervention as to the electronic media are quite powerful. Precisely because the electronic media lack the characteristics of random access and susceptibility of scanning, one who exercises the prerogative of free access intrudes himself upon an audience which can be described as captive. The audience member cannot simply flick his eye across that column or that page and get on with the balance of what that station has to offer. Thus, in the context of broadcasting, one man's freedom of access becomes another man's lack of freedom to ignore.

It seems even more clear that we should not tolerate in the electronic media both freedom of access and a continuation of the "fairness doctrine," for that combination would require the broadcaster first to provide an opportunity for the expression of idiosyncratic views and then, at his own expense in terms of opportunity cost, to present a balancing presentation. Apparently it was this prospect that led the Supreme Court to rule against the right of access in its 1972 decision involving the Democratic National Committee and the Business Executives Move for Vietnam Peace and, on the other hand, two networks and a broadcaster. Throughout its opinion the Court assumed that, if it were to recognize a constitutional right of access, the *FCC*-imposed "fairness doctrine" would nevertheless remain in place with the consequence I have already suggested (*SCR*, 1972, *93*, 2080).

In the case of the printed media, however, the reader does retain his right to ignore, and the right of access, certainly if conditioned on payment of the publisher's usual advertising rates, would appear to be a far less onerous solution than any governmental attempt to impose and administer a "fairness doctrine" on publishers.

It may be worth noting in this context that several states have long had on their statute books a highly specialized kind of "free access" provision, generally known as a "retraction" statute. These statutes, however, provide access only in a very limited sense. If a newspaper publishes a defamatory attack and if its victim demands a retraction, then the paper may, at its option, publish a retraction in its own words. In a subsequent suit by the victim for defamation, the paper is protected against certain types of damages if it has published a retraction, but it remains exposed to a wider range of damage liability if it publishes an inadequate retraction or none at all. Obviously such statutes create so limited a version of access that their history sheds little light on the behavioral changes that would be wrought by a statute imposing a general obligation to sell space, for example, at regular advertisement rates for the dissemination of social or political messages. But the long acceptance of "retraction" statutes may suggest that a broader access statute would not be held in violation of the First Amendment.

A case recently decided by the Florida courts (and not yet published by court reporting services) may prove to be more significant.[5] A Florida statute was interpreted as affording a right of reply without cost to a candidate for public office who had been criticized by the *Miami Herald.* The U.S. Supreme Court, to which the case has been appealed, will soon rule whether such a statute is consistent with the First Amendment. A decision invalidating the statute would, for the moment, perpetuate the paradox under which the many electronic voices in a locality are regulated while the few print voices are not; but it would pave the way for resolution of the paradox along traditional free speech lines by preserving the opportunity for the Court to abandon the Red Lion doctrine. To uphold the statute would be to resolve the paradox by accepting government controls over printed media. Thus, we can expect a Supreme Court decision this spring which may markedly alter the tradition of free speech in the United States.

---

[5] *Tornillo* v. *Miami Herald Publishing Co.*

REFERENCES

E. Emery, *The Press and America,* 3d ed., Englewood Cliffs 1972.

A. DeVany, R. Eckert, C. Meyers, D. O'Hara, and R. Scott, "A Property System for Market Allocation of the Electromagnetic Spectrum," *Stan. Law Rev.,* June 1969, *21,* 1499–1561.

R. Noll, M. J. Peck, and J. McGowan, *Economic Aspects of Television Regulation,* Washington 1973.

J. Rosse, "Daily Newspapers, Monopolistic Competition, and Economies of Scale," *Amer. Econ. Rev.,* May 1967, *57,* 522–533.

P. Steiner, "Program Patterns and Preferences, and the Workability of Competition in Radiobroadcasting," *Quart. J. Econ.,* May 1952, *66,* 194–223.

E. Sevareid, participant of panel discussion reported in "The Political Atmosphere," *The Center Mag.,* May–June 1973, *6,* 54–61.

Ayer Press, *1973 Ayer Directory of Publications,* Philadelphia 1973.

Code of Federal Regulation, 1967, *47,* §73.123.

Federal Communications Commission Opinions, 13, 1246.

Federal Register 1968, *33,* 5362.

Pike and Fisher Radio Regulation, 2d ser., *2,* 1901.

Statutes at Large (*STAT*), 1959, *73,* 557.

———, Act of Aug. 13, 1912, *37,* ch. 287, p. 302.

———, Act of Feb. 23, 1927, *44,* ch. 169, p. 1162.

Supreme Court Reporter, 89, 1799.

———, 1972, *93,* 2080.

Tornillo v. *Miami Herald Publishing Company.*

U.S. Code (*USC*), 1934, *47,* § 315a.

U.S. Reports, 1969, *395,* 367.

DISCUSSION

Bruce M. Owen, Stanford University: Our rebellious colonial forefathers seem to have believed that the press must be kept independent of government in order to avoid tyranny. Accordingly, the First Amendment says that the government should not regulate the marketplace of ideas. Recent Supreme Court decisions and long-standing Federal Communications Commission (*FCC*) policies have the opposite effect. The law now is that the government has a positive obligation to ensure that the public is "informed," largely by broadcasters acting under government license as fiduciaries of the public. According to the Court, there is no public right of access to the airwaves. This paternalistic attitude is at variance with the "principle" of the First Amendment, even though it is the current legal interpretation of the Amendment.

The principle of the First Amendment presumably is that there should be "freedom" (from government) in the marketplace of ideas and in the expression of ideas. A plausible corollary is that there ought also to be freedom from private restraints in this market.

There is an interesting set of connections between freedom and the economist's notion of competition. In a sense, perfect competition is the antithesis of freedom (for producers), since each economic agent is faced with a choice set containing only one alternative which permits survival. Monopolists are not constrained by this survival mechanism if they are protected by entry barriers and from take-over bids and accordingly have some freedom to make different choices consistent with survival. But the First Amendment principle is concerned not only with the production of information and ideas but also with their consumption, and here it is clear that greater competition (or at least a greater number of producers) is likely to lead to greater freedom of choice. Product differentiation in this market is not only inevitable but desirable.

The industrial organization of the mass communications media is of greater than the usual importance because of its implications for freedom of expression. Actually, we know quite a lot about media structure and technology. Most cities have only one daily newspaper firm, and this circumstance is explained by the tension between economies of scale in printing and distributing newspapers and the product differentiation which results from monopolistic competition in satisfying varied tastes and a geographically dispersed population. Most consumers have access to three or more television stations and dozens of radio stations, and this is explained by the *FCC*'s spectrum allocation policies. There are three television networks, and this is explained both by the *FCC*'s spectrum policies and by the economies of scale of distributing collective goods (programs) over large audiences. It seems generally to be true that in the *transmission* stage of most mass media there are significant economies of scale, and it also seems generally to be true that the owners of this stage (broadcast licensees, newspaper publishers) also control (or are legally compelled to control) access to their media, and thus the selection of messages. In this way, the industrial structure of the *transmission* media affects freedom of expression by producers of ideas (who are mostly atomistic) and the freedom or range of choice of consumers of ideas: the economics of this structure dictate fewness and market power for the gatekeepers of the media. By contrast, some transmission media with significant degrees of concentration—for example, the telephone company and the post office—do not control the content of messages.

The standard liberal approach to the problem of market power and fewness in the media is, first, to increase the "number of voices" by a policy of deconcentration, subsidization of new entrants, and increased access, and, second, to increase the vigor of government regulation. Thus one hears of proposals to break up joint newspaper-television interests, to bring antitrust suits against newspaper mergers and newspaper chains, to increase funding for public broadcasting, to legislate a right of access to newspapers for editorial advertisements; and to impose government restrictions on television violence, cigarette commercials, and the like. Curiously, the Nixon administration attacks on the powerful "Eastern Media Establishment" were seen as antiliberal, and it is even argued that only a large, powerful and protected press can hope to supply a viable countervailing force to government power. This poses an interesting research issue: Is a highly concentrated press better able to perform its political role than a competitive press? The answer is far from obvious. For instance, a competitive press may be "weak" in relation to government efforts to suppress dissent by lies or by direct control of information flow, but a powerful concentrated press provides a barrier to antiestablishment ideas and, more important, presents the very rationale required for government intervention, a rationale which has always been the mainstay of broadcast regulation. (There is a considerable anomaly in our present approach since we regulate the most numerous local media, broadcasters, but not the least numerous, newspapers.)

The real key to the First Amendment principle in the communication industries appears to be vertical integration. Economies of scale in transmission are antithetical to freedom of expression only if the few suppliers of this service also control access and content. When they do, control of access and content is translated into a legal, if not a logical, rationale for content regulation. In some media, for instance newspapers, it may not be possible to separate control of transmission from control of access and content. But in others, for instance cable television, such a separation may be both possible and beneficial. Giving concentrated transmission media some common carrier obligations may be a reasonable compromise between private and public restraint.

If the goal of the First Amendment really is to prevent government intervention in the press (or the information industry), and if this principle extends to the point of allowing government to intervene *structurally* so as to reduce private restraints on the competitive flow of information, then a number of quite obvious policy initiatives are available. One can begin by following Ronald Coase's suggestion that we auction off the radio spectrum. We can go on to remove the silly rules against pay-television and most of the present structure of broadcast regulation. We can decentralize or perhaps abolish public television. We can encourage the growth of cable television and be sure that there is no anticompetitive restriction of access to it. We can permit competition in the postal service. Each of these things would go some way towards increasing competition among transmitters of messages, which is the only sector in our communication system where there is any "natural" constraint on supply, competition, or freedom. Each of these things would also significantly disturb existing equities.

Competition, or at least monopolistic competition, in the production and consumption of ideas and information is plausibly consistent with freedom of expression, and arguably essential to it. Whatever the current legal meaning of the First Amendment, we are all presumably in favor of freedom of expression. It seems perfectly plain that regulation, in the words of an *FCC* judge, "is not and never can be a substitute for competition." Therefore the proper response to elements of monopoly power in the media is not regulation but increased competition. In this particular industry we ought even to consider seriously the possibility that scale economies should be given up, if necessary, in order to increase effective competition, although it is not obvious how to do this without also destroying or distorting incentives. If there is any role here for government, it is the pursuit of policies which reduce the structural barriers to increased competition in this industry. The illogic of government regulation of the media in the context of the First Amendment should not be perpetuated when there are better alternatives. The nature of these alternatives is perfectly clear. Why are they not pursued? One reason is simply that many people do not believe either in free markets or in the First Amendment. I have heard a dozen

ECONOMIC EDUCATION

Developing Independent Problem-Solving Ability in Elementary Economics

By RENDIGS FELS*

Reversing the usual order, I shall start by summarizing and then elaborate. To set priorities among the multiple objectives of the elementary course in economics, I propose putting on center stage the training of students to analyze economic policy issues for themselves. I further propose that course content be governed by usefulness for analyzing policy issues, that familiarity with the vast volume of material in typical textbooks be sacrificed in favor of mastery of a restricted number of principles and concepts, that certain economic principles now subordinated by textbooks be emphasized, and that certain other principles emphasized by textbooks be pruned. I propose to achieve the central objective by providing students with a detailed statement of high priority concepts and principles to be mastered, by emphasizing the case method of instruction, by furnishing students with a standard operating procedure for analyzing policy issues, and by requiring them to write position papers and take tests on policy issues different from those discussed in class.

I. Toward a Liberal Education

My proposed primary objective of elementary economics derives from the purpose of a liberal education. Liberal means free. A liberal education frees students' minds from the shackles of their own narrow experience. Such liberalization can be achieved by teaching students new ideas and by training them to think for themselves. College courses, including those in economics, commonly emphasize the former and neglect the latter. Everyone agrees that a student should learn how to use his head, and we all applaud when one in fact does so. But instead of training him in mental skills, we concentrate on familiarizing him with as many ideas as we can, giving him little guidance or training in the skill of using them. This is a facet of a wider phenomenon. Universities are rarely surpassed and seldom equaled at transmitting knowledge, but they are readily equaled and often surpassed at developing skills. (The chief exceptions are athletic skills, particularly football and basketball. The reason is the reward structure and motivations of coaches.) To acquire knowledge of French literature, go to college; to learn to speak French, go to Berlitz.

Why should the main objective center on economic policy rather than economic theory? In principle, a person with a liberal education might take a lively interest for the rest of his life in economic theory rather than policy, but in practice a person with one semester or one year of elementary economics will promptly forget the theory unless he has developed an interest in applying it to what he reads in newspapers and magazines.[1] For purposes of a liberal education, theory and application must go hand in hand. Learning to apply theory independently, moreover, is very difficult. We cannot assume that if students learn theory they will automatically know how to apply it.[2] It is vital for us to take on the job of helping them learn how to do it.

---

* Professor of economics, Vanderbilt University, and Secretary-Treasurer, American Economic Association. Special thanks are due the Joint Council on Economic Education for financial support of the experimental course in elementary economics on which this paper is based. Among those who made helpful comments on the paper itself are G. L. Bach, Robert L. Heilbroner, Mark R. Killingsworth, John Siegfried, and Douglas Wagner. Of the many, many people who contributed to the experimental course, it would be gross ingratitude to fail to mention Ewing P. Shahan and Robert G. Uhler.

To avoid misunderstanding, let me be explicit that theory is basic to the approach proposed here. The stress on problem solving, especially applied problem solving, is complementary to theory rather than competitive with it. Too many principles courses today are watered-down versions of intermediate theory and carry theory to the point of zero, possibly negative returns. Even students who know they will be continuing their studies beyond the principles course need instruction in what to do with it.

II. Content of the Elementary Course

To choose development of independent problem-solving skill as the central objective is a radical proposal. A majority of students would rather memorize than think. Their wishes coincide with the practice of many of their teachers. But if my proposal is accepted, certain consequences follow for the content of the elementary course. Encyclopedic coverage of the entire field of economics must be sacrificed in favor of mastery of a highly select list of concepts and principles. To use theoretical skills to solve problems requires greater mastery of the tools than is needed to answer typical examination questions testing familiarity with them. Drastic reduction of coverage would be desirable even under the conventional central objective of teaching as much theory as possible. It would be better for students to master a limited amount of theory than gain vague familiarity with all of it quite apart from what they are going to do with their knowledge, if anything. But when problem-solving skill is made the central objective, cutting down on content becomes imperative, and, as detailed below, the theoretical tools to be emphasized are changed markedly.

To say that the content of the elementary course should be reduced is trite. It is a cliché repeated endlessly and ignored endlessly. That is the way clichés arise. A truth that is universally acted on does not become a cliché. It is the obvious truth that is ignored that gets repeated *ad nauseam*. To go beyond what is trite but true, training students to deal with policy problems must mean selecting from the economist's kit those tools most useful for a liberally educated person who has studied economics for a year or less. The test of whether to include or omit a particular concept is not its importance to the professional economist nor its profundity nor its place in an esthetically elegant theoretical structure. The test is its usefulness to the layman in comparison with the time and effort he must take to learn it—to wit, its benefit-cost ratio.

Applying this benefit-cost test has led me to some familiar conclusions and to some surprises. In macroeconomics, textbooks commonly emphasize the right concepts and principles,[3] but microeconomics teaching needs a major reordering of priorities. First, a number of concepts deserve much less emphasis than they now receive. The theory of consumer choice, though it is a pillar of economic theory, is of too little use at an elementary level to warrant the time required to teach it. The basic concepts of marginalism have merit, but complex diagrams embodying the theory of the firm and of marginal productivity do not survive the cost-benefit test. On the other hand, supply and demand analysis turns out to be remarkably robust, applicable to a wide variety of problems not conforming to the strict requirements of perfect competition. Social cost, social benefit, externalities, and income distribution deserve far more emphasis than they now get. Finally, elementary texts must discuss certain issues more clearly. Economists are quite rightly efficiency nuts, but the concept of allocative efficiency must be taught at a much simpler level than that required to prove that perfect competition leads to an ideal allocation of resources. The deceptively simple distinction between current and constant dollars requires careful attention. Rationing as a function that any economic system must perform has to be separated out from the familiar what, how, and for whom.

---

[1] In his lasting-effects study, Phillip Saunders found a striking correlation between understanding of economics by alumni and the reading of certain periodicals devoted to news about business and economics. This evidence can be interpreted in several different ways, but it is at least consistent with the statement in the text.

[2] See Fels and Uhler, pp. 30–31, for an embarrassing illustration.

[3] Namely, *GNP*, the multiplier, Keynesian equilibrium, money creation, and the instruments of monetary and fiscal policy.

Students, moreover, need to be told precisely what theoretical tools they are expected to master. The extent of the material specified should be strictly limited, and students should be held to high standards of mastery.[4]

III. Instructional Approaches

The case method, which came to dominate the teaching of law and business administration decades ago, has been slow to catch on in economics. But it has great potential for our field, and in recent years there has been an upsurge of interest in it.[5] In the case method, the student is given real-world problems. He is expected to apply to these problems not only the theoretical knowledge he has acquired, but also the analytic methods characteristic of his field. Often there are no clearly right or wrong answers to the questions raised in a case: the object is to develop and to highlight the role which values and norms must play in the formulation of policy.[6] Failure to use the case method in economics in the past has meant that we have filled our students full (if not overfull) of abstract theory but have failed to train them to apply it. Even in recent years, when economics teachers have responded to students' demands for "relevance," so-called cases[7] have often been used merely to illustrate how theory is applied by economists to problems like pollution, population, the drug traffic, and prostitution, rather than to develop the student's own skill and judgment.

Why has the case method been slow to catch on in economics? The main reason has been the existence of a well-developed body of theory which teachers have wanted to teach and students have been content to memorize.[8] But to an important degree, the reason has been the difficulty of writing good cases, or rather in writing good questions and sample answers. Finding suitable material for cases in newspapers and magazines is not especially hard. The trick is to write questions that are intellectually stimulating, yet within the capabilities of beginning students. Enough progress has now been made in the technique of case preparation to warrant asserting that the central objective of the elementary course proposed at the beginning of this paper can be achieved. More precisely, it can be achieved to a sufficient degree to warrant making it the central objective.[9]

---

[4] For more precise specifications, see the Glossary in Fels and Uhler, pp. 159–169. For an earlier attempt, see Fels 1955, pp. 923–25.

Heilbroner in a letter has suggested that I "add a word about the need to instill a critical view towards the postulates of economics as another task for the introductory course. There is an awful temptation to swallow the received substance whole."

[5] A pioneering attempt was made by Aaron W. Warner and Victor R. Fuchs. More recent efforts have been made by James S. Duesenberry and Lee E. Preston; C. T. Sandford and M. S. Bradbury; Kenneth W. Clarkson and Courtenay C. Stone; G. F. Papanek, D. M. Schydlowsky, and J. J. Stern; Milton H. Spencer (pp. 296–307 and 667–77); Richard E. Attiyeh, G. L. Bach, and Keith G. Lumsden; Robert V. Horton; Donald L. Sjoquist; Lumsden; Klaus Stegemann and Walter Hettich; and Fels and Uhler.

[6] This is the sense in which the term "problem solving" is used in the title to this paper. The student approaches a policy issue with a set of value judgments and perceptions, not necessarily accurate, about the world in which he lives. Solving a problem requires that, within the constraints of what he knows, the student choose a policy option consistent with his value judgments.

[7] The word "case" has unfortunately become ambiguous, but the term "case method" is reasonably definite. (Even in law schools, however, the case method can degenerate into calling on the student to regurgitate the facts and findings of a law case instead of asking him to reason out for himself what the judge's decision should be.) Except for some of the references in the preceding footnote, the word "case" in this paper will always refer to a real-world problem with the student asked to work out the solution.

At Vanderbilt University, we have been working on the case method for several years under a grant from the Joint Council on Economic Education. Our experience indicates that two different kinds of cases are needed. The simpler kind calls for the student to use his knowledge of economic principles to interpret and evaluate an article from a newspaper or a magazine.[10] Only after students have had some training in such straightforward application are they ready for the more difficult task of thinking through their positions on policy issues. To get them beyond mere rationalization of preconceptions and to make sure they learn to approach policy issues systematically, students need to be trained to use a standard operating procedure. The procedure we teach has five steps: first, define the problem needing to be solved; second, list the main goals or objectives to be sought, the principal policy options, and the economic principles and concepts likely to be most useful in the succeeding steps; third, analyze the likely consequences of each of the policy options in turn; fourth, evaluate the options according to each of the goals in turn; and finally, come to a conclusion on the issue based not only on the evaluations but also on tradeoffs among the goals.[11]

Learning a skill requires practice. Students can be given practice in problem solving through class discussion of cases, through frequent quizzes requiring them to analyze new cases as well as ones they have already studied, and through the writing of short position papers. Alternatively, the Keller system of personalized self-paced instruction can be used (see Fred S. Keller; see also J. S. McMichael and J. R. Corey). But practice is not enough. Practice does not make perfect unless it is accompanied by quick feedback—praise for what the student does right, suggestions for how he could do better, correction of errors. Quizzes must be returned promptly with numerous comments; papers must be revised and improved in response to constructive criticism. Teaching a skill requires patience and tolerance. Learning to skate on ice means falling down; learning to solve economic problems means making blunders. The instructor in either case must be patient, tolerant, and encouraging.

---

[8] Economics teachers have correctly regarded the teaching of economic theory as their first duty. In contrast, when the case method first caught on in business schools, there was available very little in the way of principles to teach. The best that business schools could do was to make students wrestle with hundreds of genuine business problems. As social science and mathematics have come to be applied to business problems, the case method has properly been downgraded, but it remains a powerful teaching tool. Economics, in contrast, has long had a well-developed body of theory which will continue to be the central core of undergraduate courses, but it needs to be enhanced by the use of the case method.

[9] For some of the evidence, see Horton, Lumsden, and Sjoquist.

[10] For example, the student is given an excerpt from President Nixon's speech of August 15, 1971, announcing his New Economic Policy, in which he said, "Tax cuts to stimulate employment must be matched by spending cuts to restrain inflation." The student is asked if he agrees. This is a suitable question for class discussion. Enough students can answer it correctly to show that it is not too hard, but all students benefit from discussion leading to explicit explanation of the inconsistency in the speech. For further examples, see Fels and Uhler, pp. 18–33, 72–83, 113–117, and 132–35.

[11] For details, see Fels and Uhler, pp. 34–47 and 84–91. Three comments on the standard operating procedure are in order. First, since it can be used for any problem requiring decision, its value for a liberal education transcends economics. Second, training students to use it may have the undesirable side effect of leading to woodenness in their position papers and answers to examination questions. I believe, though, that the side effect is not serious. If, as is to be hoped, the students continue to approach policy issues systematically after the course is over, the woodenness will disappear. Third, the procedure has an implication for course content: the need for including explicit discussion of economic value judgments.

These how-to-do-it comments are incidental to my main theme. They have been included because selection of ends cannot be divorced entirely from availability of means. No goal, however worthy, should govern action if there are no means for achieving it. It was therefore incumbent on me to sketch a roadmap for reaching the destination. But the main thrust of this paper is to respond to R. A. Gordon's unpublished comments at last year's session on economic education. He chided the Committee on Economic Education for concentration on teaching methods and neglecting the crucial problem of course content. His criticism was just. Teachers of economics all too often proceed on the basis of tacit assumptions about objectives, assumptions that would not stand up under critical examination. I have argued that the central objective of the elementary course should be developing independent problem-solving ability. To adopt that objective as the guiding principle for course content would help eliminate the common error of overloading the elementary course and lead to the maximum contribution to a liberal education.

REFERENCES

R. E. Attiyeh, G. L. Bach, and K. G. Lumsden, *Basic Economics: Theory and Cases,* Englewood Cliffs 1973.

K. W. Clarkson and C. C. Stone, *Microeconomics in Action,* Englewood Cliffs 1971.

J. S. Duesenberry and L. E. Preston, *Cases and Problems in Economics,* Englewood Cliffs 1970.

R. Fels, "On Teaching Elementary Economics," *Amer. Econ. Rev.*, Dec. 1955, *44,* 919–32.

——— and R. G. Uhler, eds., *Caseoook of Economic Problems and Policies: Practice in Thinking,* St. Paul 1973.

R. V. Horton, "Student Applications in a Principles Course of Economic Analysis to Self-Discovered Items," Instit. for Res. in the Behavioral and Social Sciences, paper no. 354, Lafayette 1972.

F. S. Keller, "Good-by Teacher," *J. Applied Behavior Analysis,* Spring 1968, *1,* 79–89.

K. Lumsden, "The Use of Cases in Teaching Economics: An Uncontrolled Experiment in the Principles Course," mimeo., 1973.

J. S. McMichael and J. R. Corey, "Contingency Management in an Introductory Psychology Course Produces Better Learning," *J. Applied Behavioral Analysis,* Summer 1969, *2,* 79–83.

G. F. Papanek, D. M. Schydlowsky, and J. J. Stern, *Decision Making for Economic Development: Text and Cases,* Boston 1971.

C. T. Sandford and M. S. Bradbury, *Case Studies in Economics: Economic Policy,* London 1970.

P. Saunders, *The Lasting Effectiveness of Introductory Economics Courses,* Final Report, *NSF* Grant GY208, mimeo., 1973.

D. L. Sjoquist, "The Use of Case Problems in Principles of Economics," mimeo., undated.

M. H. Spencer, *Contemporary Economics,* New York 1971.

K. Stegemann and W. Hettich, *Micro Economics Applied to Press Reports,* Toronto 1973.

A. W. Warner and V. R. Fuchs, *Concepts and Cases in Economic Analysis,* New York 1958.

Computer-Aided Instruction for Large Elementary Courses

By BERNARD H. BOOMS AND D. LYNNE KALTREIDER*

Large classes are an important element of higher education in the United States today. Pressures of increased enrollment and shrinking financial support are likely to insure the continued existence of large classes in our colleges and universities, even though a sizable proportion of both students and faculty view this technique of instruction as unsatisfactory.

With increased size have come changes in the organization, atmosphere, activity, and acceptability of the classroom scene. In the typical large enrollment course, the professor meets the students *en masse* several times a week for lectures and demonstrations, and once a week the students are expected to attend small recitations sections (thirty to forty students), usually conducted by graduate assistants. The professor has very little opportunity to call on students in class or to administer weekly tests. Because of the size of the class, the monitoring of student performance is normally limited to one or two midterm examinations of one hour each in length, plus a final exam. The sheer size of these big sections influences not only the operation of the course but also the atmosphere in which it is conducted. In large classes, flexibility and degrees of freedom available to the instructor and the students are frequently lost.

Aware of this dissatisfaction with typical large class instruction, we attempted to find a system for teaching large classes which would grant maximum freedom and responsibility to the individual student; help the student to develop regular study habits, yet at the same time reduce his anxiety; give the student more than one chance to demonstrate a grasp of the material; increase his learning; and engender in him a more positive attitude toward economics in particular and university training in general.

The Computer Generated Repeatable Testing (*CGRT*) System described here is similar to that developed by Allen C. Kelley at the University of Wisconsin, although Kelley's *TIPS* program functions to a large extent as a study aid—information on the students is collected periodically, and this information is used to generate individualized assignments—while the *CGRT* System relates to the entire instructional process, including grading. This is just one of numerous new computer-assisted teaching systems being tried in many fields throughout the country. (A list of these is currently being compiled by International Business Machines. When we were contacted in 1972, the list already included forty programs.)

A computer-managed teaching system used by Donald Jensen at Indiana University in 1968 led to the *CGRT* experiment.[1] During the 1971-72 school year, a modification of Jensen's computer-managed teaching system was tested and evaluated in an introductory macroeconomics class of 320 students at The Pennsylvania State University. The class—16 percent female and 84 percent male—had an average College Entrance Examination Board-Scholastic Aptitude Test (*CEEB-SAT*) score (verbal and math combined) of 1096. The results were very encouraging.

---

* Associate professor of economics, The Pennsylvania State University, currently on leave, Economist in Residence, City Manager's Office, Tacoma, Washington; and research assistant, Institute for Research on Human Resources, The Pennsylvania State University. The original acknowledgements for this paper covered one complete page; however, the editor of this *Review* would not allow the authors to share the blame so widely. Nevertheless, two people who helped with this experiment must be mentioned—Robert Posatko and Roger Tellefsen, research assistants, Department of Economics, The Pennsylvania State University.

I. Description of the *CGRT* System

The backbone of the *CGRT* System is a testing-teaching setup that is implemented via the computer. Testing is used both for monitoring student performance and for assisting students in the learning process. Each week the instructors administer tests on a new section of material. The tests, generated via the computer and repeatable, are coupled with a tutoring service that is staffed by graduate assistants. The remaining elements of the *CGRT* System are the traditional parts of a college course—the reading material and the lectures.

The system encourages flexibility on the part of the students since they have some control over their own progress through the material and considerable discretion over when and how often their performance will be monitored. The system is powerful and "fair," in that it provides the student with more than one opportunity to be evaluated on a given body of material. The system allows students to repeat the tests with the highest grade counting without penalty. While encouraging student independence and individual responsibility by allowing students to schedule their own repeat examinations within each week, the *CGRT* System, at the same time, establishes a routine (self-paced) study pattern for students and makes demands of them. The system is personalized—one-to-one interaction is provided through a tutoring arrangement, available to all students desiring it. Finally, the computer management of the course enables course instructors to have greater time for contact with students.

The experimental course involved three regularly scheduled class periods per week. Two lecture periods plus assigned readings made up the instructional unit; the third class period each week was used for testing. If the student was unhappy with his performance on a test, he could take additional tests on the same material up to two times during the following week.

Each examination, designed to cover all the material in the instructional unit as well as some material covered earlier, was prepared by computer from a question bank of approximately 4,000 multiple-choice questions which had been selected, numbered, coded by instructional unit, and punched on computer cards. In addition, answers, comments, and textbook page references were punched on computer cards with proper identifying codes. Each examination, composed of fifteen multiple-choice questions, was a stratified random sample of the questions stored in the computer.

Repeat testing was conducted in a special testing room, open twenty-four hours a week and manned by two undergraduate assistants. As the student signed in, he received a test and an optical scanner mark-sense sheet. After completing the test and turning in the mark-sense sheet, the student received a matching answer sheet from which he could grade himself, thereby receiving immediate feedback on his performance and obtaining references to specific pages in the text and additional comments. Whenever the testing room was open, a graduate assistant was available for consultation in an adjacent room.

---

[1] The experiment itself was based on an idea developed by two Brazilian and two *U.S.* psychologists in 1964, although Jensen was apparently the first to attempt a computer-managed version of the teaching system. It was because of Jensen's interest and generous assistance that the program could be employed in economics. The *CGRT* program used was a modified version of that developed by Jensen and Franklin Prosser at Indiana University.

The completed mark-sense answer sheets were passed through an optical scanner to obtain student answers in punch card form. The cards were then compared by computer with the answer keypunch cards, and student grades were recorded. A printout of student numbers and each week's exam results were posted each week, along with a printout of the highest score for each set of tests for each student.

II. Evaluation of the *CGRT* System

Costs. It is difficult to compare accurately the costs of computer-generated individualized tests with those of conventional printed tests. However, in view of the costs of paper, printing, secretarial services, and processing, we estimate that the average cost per fifteen-item conventional test was $0.11. Our estimate of the cost of a fifteen-item computer-generated test, $0.147, was based on program read-in time, test generation costs, record keeping, mark-sense forms, optical scanner reading, and computer sorting. (See Table 1.)

TABLE 1—COST COMPARISONS

Cost Estimates Conventional Method[a]

Item		Total Cost	Cost/Test
Preparation			
Paper $1.80/1,000 sheets		$ 5.40	$.0054
Printing Supplies			
Stencils $.22/ea	.65		
Ink	.35		
Total	$1.00	1.00	.001
Secretary Services @ $3.25/hour		14.63	.0146
Typing (including rough draft)	1.5 hours		
Mimeographing	1.0 hour		
Collating	2.0 hours		
Total	4.5 hours		
Processing			
Mark-sense forms $6.46/1,000 forms		6.46	.0065
Grading			
Optical Scanner Reading		40.00	.0400
Optical Scanner Scoring		40.00	.0400
Summary Program		2.00	.0020
Total		$109.49	$.1095
Average Cost Per Test Conventional Method: $0.11			

Cost Estimates *CGRT* Method[b]

Item	Total Cost	Cost/Test
Preparation		
Computer Generation		
Program read-in time, 6 sec. @ $.12/sec.	$ .72	$.00072
Test Generation		
Ave. time/test, .400 sec. @ $.12/sec.	$ 48.00	.048
Records (lines printed)		
Ave. records/test, 80 @ $.05/100 records	40.00	.040
Processing		
Mark-sense forms	6.46	.00646
Grading		
Optical Scanner Reading	40.00	.040
Computer Scoring		
(Ave. time/test, .05 sec. @ $.12/sec.)	10.00	.010
Summary Program	2.00	.0020
Total	$147.18	$.14718
Average Cost Per Test *CGRT* Method: $0.147		

[a] Based on 1,000 three-page tests.

[b] Based on 1,000 three-page tests of the conventional type. The initial, one time, start-up costs of the *CGRT* System are not included in these calculations.

There appears to be little hope that the *CGRT* System will lead to lower total testing costs both because the System entails a higher estimated per unit cost and also because with the *CGRT* System, the *amount* of testing is much greater.

The student viewpoint—immediate. An effort was made to evaluate the *CGRT* System both objectively and subjectively. The Test of Understanding College Economics (*TUCE*) was administered at the end of the first week of the course and again at the end of the term, to evaluate objectively how much students had learned in the course using the *CGRT* System and how their performance compared with that of other students throughout the United States who had completed a similar elementary economics course. Both times the test was administered to all students simultaneously. The *CGRT* students showed significantly more gain ($p<.001$) than the students in the normative sample (see Table 2). Specifically, *the CGRT students did 23.5 percent better than national norm students.* The percent gain for the *CGRT* students was 50.4; the percent gain for the norm was 40.8—a percentage point different of 9.6 points, or 23.5 percent better for the *CGRT* students.

TABLE 2—TUCE SCORE COMPARISONS
(Forms combined)

	Normative Data[a]		PSU Economics (*CGRT*)	
	Mean Score	S.D.	Mean Score	S.D.
Pretest	13.52	4.60	14.61	4.12
Posttest	19.04	5.13	21.96	4.54
Gain Score	5.52	5.03	7.36	4.41
Pre-Post Correlation	.47		.49	
Number of Students	957		320	

t-test for Gain Score for Normative Data and *PSU* (*CGRT*): t value$=5.85$ ($p<.001$)

[a] Derived from Table 8, *Manual for TUCE: Part I—Means, Standard Deviations, and Correlations by Sequence of Form Administration,* Joint Council on Economic Education, New York 1968, p. 18.

The subjective evaluation was carried out via two student opinion questionnaires—the first, administered in class in the middle of the term; the other, during the final week of class. Anonymity for respondents was guaranteed. The following elements were considered in the evaluation: (a) course format, (b) instruction, (c) *CGRT* testing system, (d) tutoring arrangement, and (e) general reactions. The results indicated that:

(1) 85 percent of the *CGRT* students would recommend the course to fellow students;
(2) 82.3 percent felt that the computer-generated format was a good or excellent method of college teaching;
(3) 66.4 percent felt the way the course was managed provided them with extra motivation and active interest in economics;
(4) 98.9 percent of the *CGRT* students judged the immediate feedback feature on tests to be very helpful;
(5) 90 percent of the *CGRT* students felt the repeatable tests provided a worthwhile learning experience;
(6) 89 percent found the tutoring service helpful, and 73.6 percent preferred it to the "conventional method" of small group recitation sessions with graduate students;
(7) 43.7 percent felt that the *CGRT* economics course was one of the most interesting, informative, useful, and personally helpful courses they had taken; and
(8) 69.7 percent felt that the course would be of considerable value after college.

The student viewpoint—lasting effects and impressions. In an effort to evaluate the lasting effects of the *CGRT* System, we conducted a retake survey one year after the course's conclusion. Since, to our knowledge, no strictly comparable study has been made of the lasting effects of another method of instruction, the results are merely reported here, without conclusions being drawn.

The retake survey was composed of two parts: the thirty-three-item *TUCE* given a year earlier was readministered, along with an attitudinal survey of the *CGRT* System. Seventy-three of the 320 *CGRT* students took part. *CEEB-SAT* scores were used to compare means. The null hypothesis is that the mean of the retake sample is the same as that of the population. The results indicate that the null hypothesis cannot be rejected at the .01 significance level. (*CEEB-SAT* scores were unavailable for thirty students in the class population.) See Table 3. In addition, the distributions of course grades for the entire class and for the retake sample were comparable, within the limits of the population available. It was found that after a year the pretest to posttest gain achieved during the course (6.87 points) had been eroded by 3.91 points, or 56.8 percent of the initial gain on the *TUCE* (see Table 3).

TABLE 3—RETAKE RESULTS

	Class Population		Retake Sample	
	Mean Score	S.D.	Mean Score	S.D.
	$N=320$		$N=73$	
Pretest	14.61	4.11	15.54	4.06
Posttest	21.96	4.53	22.41	4.25
Retake test	—	—	18.50	5.59
	$N=290$		$N=73$	
CEEB-SAT verbal	518	81.99	533	81.18
CEEB-SAT math	578	88.74	599	85.75
CEEB-SAT total	1096	143.21	1133	134.15

Z-test for *CEEB-SAT* verbal: $Z=1.69$
Z-test for *CEEB-SAT* math: $Z=2.01$
Z-test for *CEEB-SAT* total: $Z=2.18$

Results of the attitudinal survey were interesting because of the high level of positive feedback; for example, 87 percent of those with a year of additional coursework since the initial exposure to the *CGRT* System felt the *CGRT* System was a good or excellent method of college teaching. More work is currently being done to analyze the lasting effects data, including the testing of alternative models of explanation of the results.

The graduate assistant viewpoint. The *CGRT* graduate assistants—all of whom had previously been instructors in recitation sections under the conventional class method—generally seemed to feel that:

(1) the tutoring experience was more rewarding and beneficial to them than conducting recitation sections;
(2) they worked harder under the *CGRT* System (due in part to start-up difficulties with the system);
(3) They learned more because, to be able to discuss questions directed at them by students, they prepared class work more carefully;
(4) the tutoring system freed them from much of the drudgery usually associated with the grading of midterm and final examinations under the conventional system; and
(5) meeting with students on a one-to-one basis seemed to give them a greater sense of the personal element.

III. Conclusions

The *CGRT* System seems to hold great potential for use in a wide variety of situations since it appears to be readily adaptable to individual needs and specific desires. For example, it could be modified (1) to make it totally student-paced; (2) to extend it to supplemental material, so that additional units of credit could be offered to students wishing to explore a subject further; (3) to strengthen the background of undergraduates headed for graduate school by having them serve as assistants in small discussion sections.

The results of a one-term experiment are not the last word. Final conclusions await further study. Preliminary results, however, are encouraging. The *CGRT* System appears to reduce aversion to large classes and increase the learning that takes place in such a setting. In addition, the system can be viewed as an attempt to raise productivity in institutions of higher learning, an issue of crucial importance to the future of higher education.

REFERENCES

D. Jensen and F. Prosser, *Conference Proceedings,* American Federation of Information Processing Societies, Spring 1971, *38,* 295–301.

A. C. Kelley, "An Experiment with TIPS: A Computer-Aided Instructional System for Undergraduate Education," *Amer. Econ. Rev., Proc.,* May 1968, *58,* 446–457.

Joint Council on Economic Education, *Manual for TUCE: Part I,* New York 1968.

Introducing Freshmen to the Social System

By KENNETH AND ELISE BOULDING*

Under a grant from the Joint Council on Economic Education we conducted a course for freshmen entitled "An Introduction to the Social System" at the University of Colorado in the spring semesters of 1972 and 1973.[1] We are not sure that it deserves to be called an "experiment," for unlike many of the experiments in the teaching of economics, this experience had no control groups and we did not attempt analytically to differentiate it from others. It was more like the making of a work of art than an experiment. There seems to be no reason, however, why a course should not be treated as a work of art, defined perhaps as a structure designed to enhance the consciousness of those in its environment.

The particular form of enhancement which this course was designed to produce for freshmen was a greater understanding of the social system of the world in its totality. We hope the students can learn to perceive the social system, or the sociosphere, as indeed a part of the total system of the world, having some coherence and order of its own and different in a good many respects from the other systems of the world—the biosphere, the atmosphere, the lithosphere and the hydrosphere. In a single semester it is naturally impossible to spell out this order in any detail. If, however, students as a result of this experience come to realize that the social system has degrees of order in it and is at least a partial unity, it is hoped then that they will be stimulated to pursue their studies in the particular disciplines, guided by a sense of filling in particular details of a larger picture rather than of experiencing isolated and unrelated disciplinary systems. Nevertheless, there is an inevitable paradox in this procedure, that the whole is in fact made up of parts and the system is composed of details. Without some sense of detail there can be no vision of the whole, and if windy generalities and vague formulations are to be avoided, there must be awareness of the "minute particulars" of which the world is composed, as well as the great frameworks within which these particulars are set.

The thread which held the course together, often perhaps a little tenuously, was a science fiction story, written by the authors with the assistance of their son Russell Boulding. The story is briefly that earth's broadcasts, which now extend into space for about 65 light years of structured radio waves in all directions, finally have been perceived by intelligent beings with a very high level of technology, in a satellite of Alpha Leonis. Alarmed by these manifestations which seem to indicate a dangerous knowledge explosion, they send in a scout named *EXOC* to investigate. He comes in towards earth at the speed of light, monitoring all its radio waves—of telegraphy, radio and television—and lands on the moon at about the time our class begins. He would, of course, oxidize if he came into the earth's atmosphere. He then establishes extrasensory communication with the minds of the two instructors, and under his inquiries and directions the instructors then turn the class into a research institute to inform him about the earth.

---

* Institute of Behavioral Science, University of Colorado.

[1] A complete text and syllabus of the course in mimeograph form is available from us for a fee of $10.

The dramatic device serves, we hope, to stretch the students' imaginations and to bring home two very important concepts right away. The first is the concept of the earth as a total system, producing certain outputs in the form of structured radio waves, which are almost exclusively a product of the sociosphere. In the absence of the human race, the earth can produce nothing much but snaps, crackles, and pops from occasional thunderstorms or other electrical disturbances. Now suddenly earth's output of radio waves is loaded with information, suggesting the development of structures far beyond the probabilities of the simple biosphere. The first task of the class indeed is to draw a map of the world as our extraterrestrial visitor sees it, with about 6,000 radio and television stations in North America, only about 300 in Africa, and varying densities in other parts of the world.

The second concept which is brought out right away is that of the sample. Radio and television broadcasts are obviously a rather peculiar and highly biased sample of the total output of information from the earth's sociosphere. As a result, we suppose that our visitor receives initially some quite startling misconceptions, especially as he runs into phenomena with which he is unfamiliar, such as death, he and his kind being virtually immortal. We invite the class, therefore, to struggle with the problem of what the visitor would learn from earth's broadcasts, what misconceptions he would form, and also what correct perceptions he may form. We then go on to point out that the bias in our visitor's sampling of information is probably no greater than any single human being experiences in the course of a lifetime, and this raises immediately the dangers of generalizing from a personal experience, which is a very imperfect sample of the whole, to the totality. The object of the course is certainly not in any sense to degrade the importance of the student's personal experience for the self, or indeed to deny its reality. It is, however, to show that personal experience is a very limited sample and that the need for a social science arises precisely because of the sampling limitations of the personal experience.

This opens up immediately the question of how can we derive knowledge about very large systems which are beyond our personal experience, and this introduces us immediately to the concepts of data and data collection and analysis. It introduces us also to the concept of quantification and of indicators. In trying to explain the obsession of the human race, or at least of social scientists, with numbers, indicators and statistics, we find we have to interpret this in terms of attempts to overcome certain deficiencies of the human nervous system, especially its inability to perceive directly the properties and structures of large sets. *EXOC*, of course, can perceive large *n*-dimensional sets of great heterogeneity immediately. He naturally thinks numbers are the least interesting properties of sets, but he forces himself to perceive the world through the limitations imposed by the structures of the human nervous systems with which he is in contact. This we hope gives the students a sense both of the necessity and also of the limitations of data collection and quantitative analysis.

From this introduction, the course proceeds through a wide variety of exercises, each taking about a week, which introduce the students to problems of information collection and analysis. Ideally there should be a "bank" of these modules out of which each individual instructor can select those that he prefers. So far, we have built up about enough of these for a single semester. One of these is a "random history project." Having monitored only about 60 years of human history, *EXOC* naturally wants to find out something about the total dynamic process which has produced the earth as it exists today. A whole curriculum could be built around this project in geology and biology, but time limitations force us to confine ourselves to human history. Ideally we should sample the approximately 60 billion human beings who have ever lived, but this is difficult and we compromise by confining the sample to about the last 4,000 years. Into one hat, therefore, we put cards showing the decades of the last 4,000 years, the number of cards approximately in proportion to the human population at the time, so that there are more cards for the more recent decades with larger populations. In another hat, we put cards giving a set of regions of the world, divided so that approximately equal numbers of people have lived in each region. Each student then draws a card from each hat, so that he gets a region and a decade. We then ask the student to imagine just having stepped out of a time machine, shall we say, in Cambodia in 810 B.C., and say what it is like. Each student then has to go to the library and try to find out what the region he has picked was like in the decade he has picked, provided with a schema to guide the search. After these have all been turned in we analyze them. In spite of the smallness and extreme imperfection in the sample, we hope that the student at least gets some idea as to what the total experience of the human race has been like.

The first year we did this, the class was a little shocked that American history was not even drawn in the sample, which is not surprising in view of the fact that Americans do not comprise more than five thousandths of the human beings who have ever lived. This statistic may give us at least a slight sense of humility and proportion. The approach raises some very interesting questions about how history is written and interpreted, and about the value weighting in different parts of the total human race. We do not defend the proposition, of course, that all written history should be a random sample of the total human race, but if it is not random the question of the principles of this nonrandomness at least becomes very interesting.

We have done a rather similar project on "world heroes." Our visitor is interested in the peaks of human history, as well as its averages and aggregates. He is particularly interested, therefore, in those people who are regarded as the exemplars and heroes of the human race, and we ask the class to nominate these. Then we fit these into the sample of human history as we have drawn it. Another module is the development of a tabular history of the human race by regions, which is intended to give the student a broad overview of the variety and intricacy of the human experience as it is spread out in time and space.

From this very broad overview of the space-time pattern of the sociosphere, we then turn to examine in much greater detail the sociosphere as it exists at a moment of time, namely, the approximate present. In order to do this, we have developed an extensive data book, giving the values of some sixty variables—economic, social, demographic, etc.—for all the countries and territories in the world, as of about the last ten years. Thus, from the "movie" of history, we turn to the flashlight photograph of the world as it is today. It is a little unfortunate that we have to give the students this data, except that there is no way they can go out and collect it. We do think, however, that it is important for the student to realize that data of this kind is available and that any overall picture of the world must be built up from it. Each student is assigned a variable, or two, depending on the size of the class, to be studied and analyzed, and is also given a country or region in which to become expert. There is enough material in these for a great deal of analysis.

We do not develop any elaborate statistical techniques. This would be the task of later courses, but we do suggest the more obvious things that can be done with numbers, including mapping, ranking, averages of various kinds, ratios of various kinds, and finally a very elementary form of correlation involving 2×2 matrices. Each student selects 2 variables, each of which he ranks as either "high" or "low," and divides the countries and territories into the 4 categories high-high, high-low, low-high, and low-low. A linear relationship, if it exists, will emerge out of inspection of these matrices. Clustering in the low-low and high-high boxes reflects a positive relationship; clustering in the high-low and the low-high boxes a negative relationship. The absence of clustering implies no relationship. This may be extended to 3×3 or 4×4 matrices, from which nonlinear relationships can easily be detected. One of the most interesting relationships to emerge, incidentally, is that between *GNP* per capita and latitude, the rich having a remarkable tendency to inhabit the Temperate Zone and the poor to inhabit the tropics.

After several weeks of this cross-sectional analysis of the planet as a total system, we move to a more intensive analysis of the United States, which is after all the area in which the *EXOC*-informing enterprise is going on. To get a sense of the "local dynamics" of the twentieth century, we provide some modules on the analysis of quantitative time series, both with sociological and economic data, using Bureau of Census data from 1900 on and elementary national income statistics from 1929 on. When these time series data can be presented in terms of proportions of a totality, or "layer cake diagrams," they can be extremely helpful in giving the student a feeling for the total time patterns of a complex structure. National income statistics are particularly useful in this regard, as the gross national product (*GNP*), or better, a concept named the gross capacity product (roughly equal to what the *GNP* would have been without unemployment), can be broken down into its components, such as household consumption, gross private domestic investment, government consumption, and so on. National income can also be broken down into its distributional components. The study of the fluctuations of these can be extremely suggestive. Time series using sociological data does not delineate trends with the same kind of precision, but a picture of improvements in family life style, in health, social and cultural resources and in the total knowledge stock, can be set over against a picture of increasing personal, social and economic costs generated by the phenomenon of urbanization. This type of time series data sharpens the questions about the nature of social and economic progress in the twentieth century.

Another module using economic data involves an exercise with the balance of payments matrix, showing the inpayments and outpayments from three parties, with the resulting balance and the changes in the money stock, a series of transformation rules to transform the matrix from "one week" into that of the next, and so on, for several weeks. The object of this exercise is to give the students a feeling for the peculiar properties of closed systems, whereby the effects of the decision on one person depend on the decisions of others. The students also get insights on how most systems tend toward equilibrium, though there was one interesting case in which the system turned out to be explosive. And the students get the feeling, one hopes, that the difference between systems which move towards equilibrium and those which are explosive may be a very small difference in the transformation rules.

One of the important goals of the course is to enable students at one and the same time to develop a feeling for the globe as a total system in time and space, and yet to be able to locate their own society and themselves as individuals in that system. In addition to the data and exercises provided for the United States, one exercise is designed to help locate the student in the United States time series. A family tree questionnaire is administered to generate data for the class on a three-generational experience of mobility in relation to rural-urban residence, occupation, education, religion, family size and sex roles.

The time series data provide a transition from the use of the tools of economic analysis to the tools of sociological analysis. Using the device of emphasizing *EXOC*'s hive-like social existence in a near-immortal and conflict-free society on his home planet as a foil, the emergence of stratification systems on earth is presented as a device for dealing with short-lived human groups scrambling for scarce resources in a limited environment. The structure-creating aspects of human behavior become noticeable precisely because of the short life span of human populations. Processes such as social bonding become crucial, and *EXOC* gets involved in examining learning theory and socialization processes to see how humans can function without a "hive."

At the same time that students are getting an extraterrestrial view of social processes they have always taken for granted, the simulation game *SIMSOC* is introduced. An entire day is set aside for the game, so there is total immersion in the process of creating a new society. The conflicts that arise between regions in the society as individuals and groups try to deal with resource scarcity, severely limited communications capabilities and stresses caused by imposed "natural catastrophes," lead to a great many insights into the nature of social processes, and political and economic power as altruism and exploitation, cooperation and competition, trust and mistrust, flourish side by side. Among other things, the interaction between behavior and structure becomes clear. The simulation thus becomes another means for placing the individual in the macro-system.

At the end of the semester, as *EXOC* prepares to head for his home planet, we try to bring out the intersection of the micro and the macro in the social process and to sum up the course in some tentative projections about the future of the planet from *EXOC*'s point of view.

One way to describe what we have done is to say that we have attempted to provide a markedly novel approach to the problem of overall mapping of the social terrain. The course was designed for freshmen because we felt that with this general systems type approach for a starter, the students could make much more creative use of the other information they would gain later. As with any attempt to experiment with new ways of presenting knowledge, there were problems. While we were both committed to the macrosystem approach to the study of the planet, we differed in the weight we gave to the importance of introducing the individual person into the social system, and to the experiential mode of learning as contrasted to the formal cognitive. And we both found difficulty in developing formats for the assignments that would give students the same feeling of involvement and excitement that we ourselves felt. While nearly everyone found the time-space sampling exciting, there were complaints about too many trivial exercises.

None of these problems is insuperable. One way to make the transitions in level and mode of learning more appealing to the students would be to invite a group of seniors from economics, sociology and political science to take this course as a project in preparing freshmen for understanding social science approaches to mapping social reality. The semester could then be spent in reworking these materials in ways that will enhance and perhaps expand the original purposes of the course.

REFERENCE

W. Gamson, *SIMSOC: Simulated Society*, 2d ed., New York 1972.

ENDOWMENT INCOME, CAPITAL GAINS AND INFLATION ACCOUNTING

The Investment Income Formula of the American Economic Association

By DONALD A. NICHOLS*

Economic theory tells us that there is no single "best" way to treat capital gains in a definition of income. But it is necessary for practical reasons that some policy toward capital gains be adopted by institutions with endowments. This creates a dilemma which can be resolved only by a careful blend of theoretical analysis and judgment. I will attempt to clarify the roles of theory and judgment in this paper, and then to outline their implications for the income formula of the American Economic Association (*AEA*).

I. Intertemporal Resource Allocation

The theoretical aspects of the capital gains accounting problem can be illustrated most easily in the framework first developed by Irving Fisher. Fisher describes how a consumer who receives satisfaction from consumption both in the present and in the future maximizes his utility. In the two-period example of Figure 1, the utility derived from present and future consumption goods (C_0 and C_1) is represented by the indifference curve I. The budget constraint is a list of the receipts expected in each time period (R_0 and R_1) and the interest rate (i) tells the price at which these receipts can be traded for each other. The present market value of the portfolio of expected receipts is represented by V.

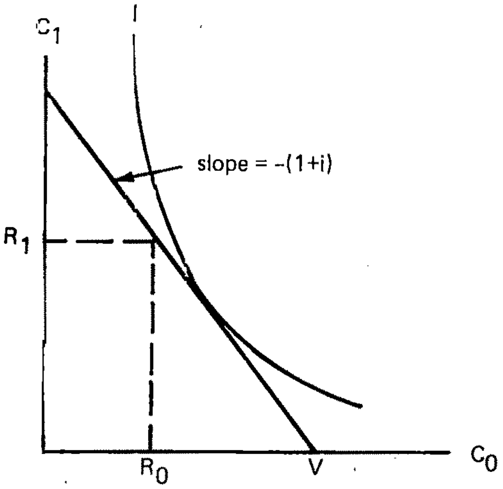


FIGURE 1

Capital gains occur when V increases in value. But V can increase as a result of several things. Figure 2 illustrates capital gains of equivalent size that result from different causes. V can increase to V^* either because of an increase in R_1 to R_1^* or because of a reduction in i to i_1^*. These two different kinds of capital gains move the utility-maximizing consumer to different levels of welfare (i.e., from indifference curve I to curve II or III), and they cause different levels of present consumption to

---

* University of Wisconsin. I wish to thank James Tobin, Henry Wallich and William D. Nordhaus for their comments on an earlier draft of a different paper on a similar subject. This research was supported at one time or another by the National Science Foundation and the Institute for Research on Poverty at the University of Wisconsin.

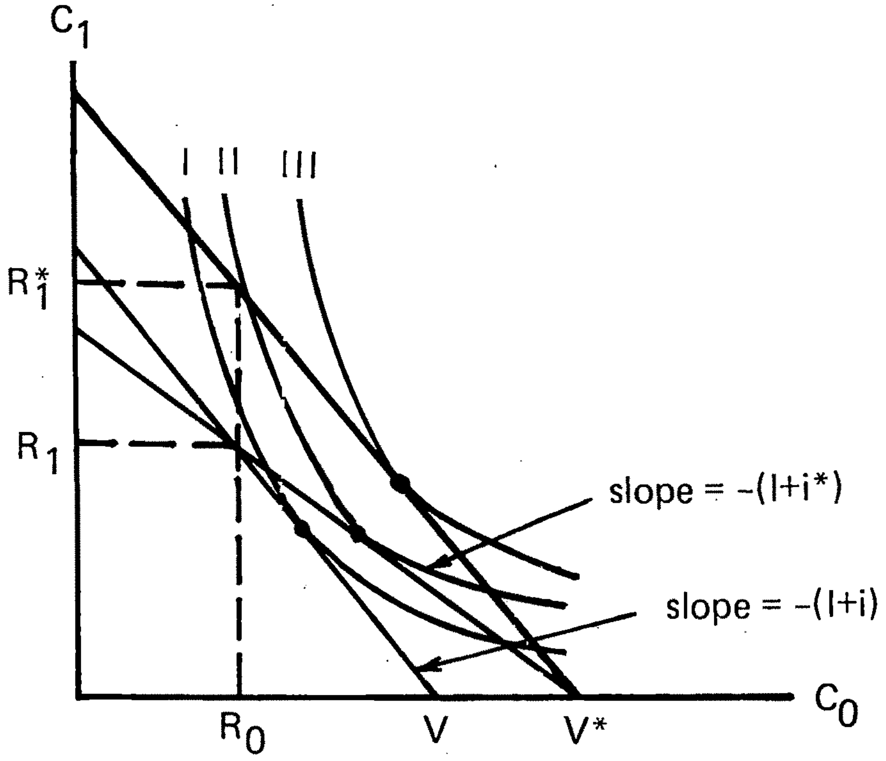


FIGURE 2

be chosen. While it is difficult in practice to be sure that a perceived gain is of one kind or the other, it is necessary to treat them differently if utility is to be maximized.

II. The Role of Income

At this point, the question of how to define income arises. It should be noted that we have been able completely to specify the investor's intertemporal decision problem without mentioning income. There is a utility function and a budget constraint, and by maximizing one subject to the other, the optimal level of present consumption can be chosen. Income is not needed. The role that income usually plays can best be seen by breaking the maximizing decision into two parts: (1) certain characteristics of the budget constraint are summarized by a number called income, and (2) consumption is determined given that level of income. This two step process is inefficient in the sense that no single number can contain all the information necessary to attain the maximum level of utility. But it may be convenient in that it allows the decision maker to sidestep a host of knotty theoretical and empirical problems, and thereby economize on decision-making costs. This seems to be the view of John R. Hicks:[1]

> In spite of its familiarity, I do not believe that income is a suitable tool for any analysis which aims at logical precision. There is far too much equivocation in its meaning, equivocation which

---

[1] The original passage referred to income as one in a list of concepts. Accordingly, I changed all plural references to singular.

> cannot be removed by the most painstaking effort. At bottom, it is not a logical category at all; it is a rough approximation used by the business man to steer himself through the bewildering changes of situations which confront him. For this purpose, a strict logical category is not what is needed; something rougher is actually better. [p. 171]

The "rough approximation" would be better if it was so much easier to use that the reduction in decision-making costs compensated for the loss due to the imprecision of the solution.

III. Hicksian Income

Hicks is often cited by those who favor the inclusion of all capital gains as a part of income. According to Hicks:

> The purpose of income calculations in practical affairs is to give people an indication of the amount which they can consume without impoverishing themselves. Following out this idea, it would seem that we ought to define a man's income as the maximum value which he can consume during a week, and still expect to be as well off at the end of the week as he was at the beginning. [p. 172]

I will interpret being "as well off at the end of the week as he was at the beginning" as meaning that the expected receipts and intertemporal prices which prevail at the end of the week will permit the purchase of the exact bundles of future consumption goods that were planned at the beginning of the week.

Contrary to what is commonly asserted, this Hicksian definition does not define all capital gains as income. As can be seen in Figure 3, the two kinds of capital gains lead to two different levels of income, Y_0, where income is defined as the most that can be consumed today while leaving future consumption (C_1^*) unchanged. Capital gains of equivalent amounts should affect income in different ways according to Hicks.

As just described, this measure of income does not economize on required information in the manner that is necessary if income is to serve its purpose. One would still have to know all intertemporal prices, all expected receipts, and the previously planned levels of consumption for each period. Not only does this make the calculation of income a tedious job, but if this information is truly available, it is probably more efficient to maximize utility directly without calculating any intermedi-

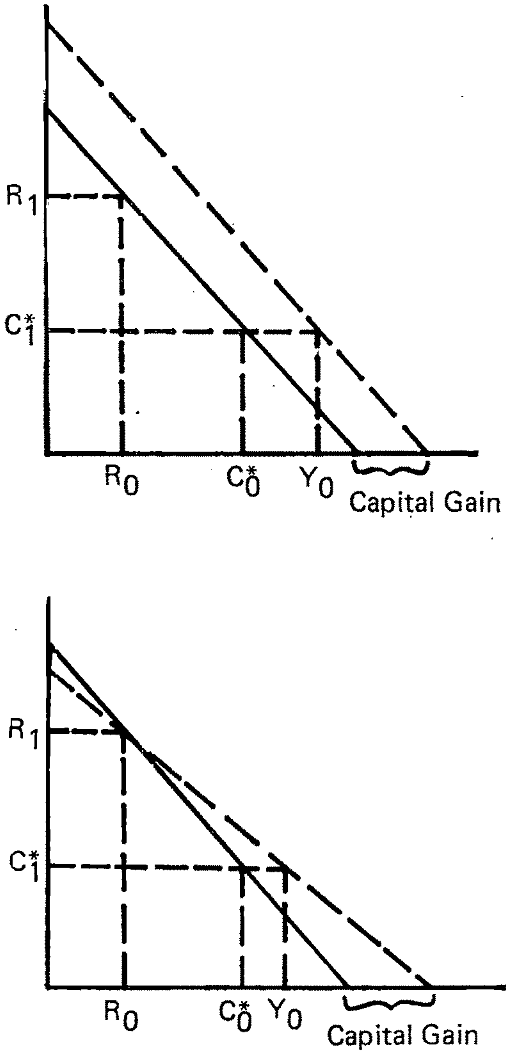


FIGURE 3

ate approximation called income. The only approximation in this case is to the utility function, not to the budget constraint, and its use requires the decision maker to throw away useful information.

IV. The Importance of the Consumption Plan

It is important to note that Hicksian income varies with the consumption plan. One can't determine Hicksian income in the absence of some knowledge of the utility function. This knowledge is necessary when intertemporal prices change. If expected receipts stay constant while intertemporal prices change, it is important to know whether the asset holder had intended to buy or sell those commodities which have become relatively more expensive before we know whether he has been made better off by the price change. This is illustrated in Figure 4, where two consumption plans are denoted by the two rays from the origin *M* and *N*. These are plans for consumption of present and future goods in fixed proportion. Plan *M* involves a great deal of future consumption, which involves the use of some present receipts to buy future consumption goods. Clearly, prospects for this plan are made worse off when interest rates fall (and future goods become more expensive). Prospects for plan *N* improve in that case, however, because plan *N* requires the sale of some expected future receipts to finance present consumption. The more valuable are those receipts, the more present consumption they will bring.

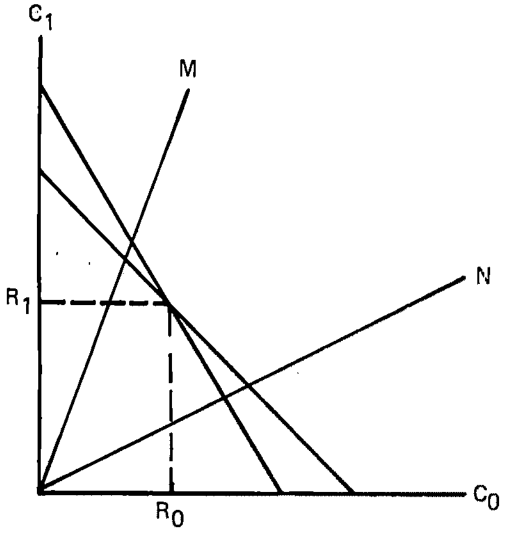


FIGURE 4

It should be obvious that two individuals (or institutions) that expect to follow plans *M* and *N*, respectively, should not use the same definition of income. The change that has just been described in the market value of the portfolio should be a signal to *N* to increase consumption and to *M* to reduce it. Clearly, the choice of an income definition should not be made independently of some knowledge of the utility function. Income is therefore subjective, and the choice of one definition over another requires some judgment on the part of the recipient concerning his future plans. The effect of this judgment is to determine how the two different kinds of gains are to be treated.

V. Perpetual Income

While no income definition can claim general usefulness, the consumption plans of many institutions are similar enough that a single definition can serve them all. Perhaps the most common plan is one in which consumption grows at a constant—sometimes zero—geometric rate. This, I would argue, closely approximates the plans of many universities, hospitals or galleries with endowments and is a good first approximation to the needs of the *AEA*.

Elsewhere I have derived a definition of income that is appropriate for this case if expected receipts from the portfolio can also be characterized by an exponential function. In this case, the income shown in (1) is exactly equal to the utility-maximizing level of present consumption.

$$\text{(1)} \quad \text{Income} = R + (G_R - G_C)V.$$

R, receipts, is the rate at which dividends and interest are being earned on the portfolio, G_R and G_C are the geometric rates of growth of receipts and consumption, and V is the present market value of the portfolio. As before, it is true that an increase in V caused by a change in discount rates—changes not accompanied by changes in R or G_R—should lead to an increase in present consumption only if the institution expects to be a net seller of securities. Obviously, $(G_R - G_C)V$ must be sold each time period.

Whether the advantages and disadvantages of my definition make it better than some other in a particular instance is a matter of judgment. If the choice seems difficult, perhaps some third measure should be used, or perhaps both. Since we live in a calculating age, it is probably no great burden for present administrators to be given two (or more) measures of income to guide them in their consumption decisions. One, dividends plus capital gains, could be used to tell them how fast their ability to buy present goods is growing, while the other, my definition, tells them how much they can expect to spend from the endowment forever. For the technically adept, income can be dispensed with entirely and the utility maximization problem can be solved completely. The whole problem of approximation, including that of choosing a treatment for capital gains, will not exist in that case.

VI. Principles Affecting the *AEA*'s Choice of an Income Definition

Because the income from the *AEA* endowment is very small relative to expenditures, the choice of an income formula is not important in the sense that even a very misleading definition will have little effect on our intertemporal spending policy. I am told, however, that the formula is important since it will serve as an example to other institutions for whom the choice is of more fundamental importance.

If it is true that the definitions used by economists have some influence in the practical world, then we have done a great disservice in the past by our continual adoption of dividends plus capital gains (Haig-Simons income) in our research and in our past recommendations to educational institutions as a good practical measure of income. Empirical research on portfolio or consumption decisions should at least try several different income definitions to show our awareness of the arbitrary nature of the income concept.

If income is subjective, as I am sure it must be, then there is some question whether the *AEA* can set an example. Only if our intertemporal utility function coincides with someone else's should he adopt our chosen definition.

Because the *AEA* endowment is small relative to annual expenditures, its value can be affected substantially by small changes in our costs or income. The endowment acts as a buffer or residual to the budget rather than being a major determinant of the size of the budget. It is an inventory of purchasing power—like a consumer's savings account—which smooths out fluctuations in income and spending. For this reason, the value of our endowment is of importance to us in a way that it should not be to an institution whose endowment is a good deal larger than its annual expenditures. Since capital losses will constrain *AEA* financial freedom, they should be replaced, and in this sense, all gains should be counted as income for us where they shouldn't necessarily be for others.

On the other hand, we are an intermediary. We have no utility function other than those of our members. If we were to perform the Fisherian comparison of the maturities of our needs and receipts, it should be the receipts of our members that are the relevant magnitudes. The endowment is like an economists' mutual fund whose dividends are received in the form of lower dues. This makes us differ from any institution that does not have a similar taxing power. One indication of this power is the fact that we decide how large our dues will be. Thus our income is a matter of choice, not a result of a fixed constraint. Given an income and spending policy, the policy toward the endowment merely tells us whether our dues should be high now or in the future.

If this last principle is used to guide investment decisions as well, it tells us to choose investments whose returns are negatively correlated with the rest of the returns in economists' portfolios. Since the largest part of those portfolios is in the form of economists' income, this suggests that we should own stocks which are apt to perform well when professional salaries perform badly and vice versa.

Finally, we should note that the *AEA* endowment could be used to carry out a small program of intergenerational transfers. Because of the continual change in *AEA* membership, the endowment resembles a government trust fund more closely than it does a private mutual fund. If it was thought that a constant level of research was better in the long run than one that fluctuated dramatically over long periods of time, we could build up the *AEA* endowment during periods such as the present when research money is plentiful and draw it down in the future if the times change. A similar policy could be followed with respect to our salary changes. Dues could be kept constant when salaries remain constant and increased substantially when salaries are growing. The endowment could grow and shrink in response to those needs, while the welfare of economists grows smoothly. If the endowment is to serve such needs, then we must be concerned with its principal value since net sales are in prospect at some date and sellers of securities always profit from high prices.

For all these reasons, the *AEA*'s endowment could be treated differently from that of a private university. Since there are many kinds of nonprofit institutions with endowments, it would be instructive to debate the merits of each of these questions in print so that different institutions might choose different income formulae for the right reasons.

VII. The *AEA* Formula

Having gone this far in outlining the impossibility of creating an objective measure of income which gives correct spending signals and having pointed out the many different ways in which our endowment might be used, I must now address the problem I was assigned and choose a definition of income from endowment for the *AEA*. This, as I have warned you, involves the exercise of personal judgment rather than analytic powers.

Of all the considerations mentioned above, two dominate. First, our endowment income is small relative to our budget and trivial relative to the combined incomes of our members, for whom we are just an agent. For this reason, our choice of definition will have almost no effect on our members' utilities and therefore is a matter of little importance. On the other hand, we can use this opportunity to set an example, both to other institutions with endowments and to the accounting profession in general. Since many of these institutions have perpetual consumption needs, the best choice would seem to be the definition I derived for that case. Since the market values of our portfolios for this and last year are reported anyway, there is little additional information yielded by tacking accrued capital gains on to receipts and calling the sum income. Haig-Simons income conveys less additional information than does perpetual, or perpetually growing income.

In any event, perpetual income is a very useful concept for many purposes, and it deserves greater use. We could encourage that use by adopting it as our own measure of income. Income, after all, is what we think it is, not what some formula tells us.

REFERENCES

I. Fisher, *The Theory of Interest,* New York 1930.

R. M. Haig, "The Concept of Income: Economic and Legal Aspects," in R. Musgrave and C. Sharp, eds. *AEA Readings in the Economics of Taxation,* Homewood 1959.

J. R. Hicks, *Value and Capital,* 2d ed., London 1964.

D. A. Nichols, "Measuring Income From Endowment," mimeo., 1973.

H. C. Simons, *Personal Income Taxation,* Chicago 1938.

Ford Foundation Advisory Committee on Endowment Management, *Managing Educational Endowments,* New York 1969.

What Is Permanent Endowment Income?

By JAMES TOBIN*

The trustees of an endowed institution are the guardians of the future against the claims of the present. Their task is to preserve equity among generations. The trustees of an endowed university like my own assume the institution to be immortal. They want to know, therefore, the rate of consumption from endowment which can be sustained indefinitely. Sustainable consumption is their conception of permanent endowment income. In formal terms, the trustees are supposed to have a zero subjective rate of time preference.

Consuming endowment income so defined means in principle that the existing endowment can continue to support the same set of activities that it is now supporting. This rule says that current consumption should not benefit from the prospects of future gifts to endowment. Sustainable consumption rises to encompass an enlarged scope of activities when, but not before, capital gifts enlarge the endowment.

The university has other sources of income, and one might expect its permanent income, its sustainable consumption, to be defined with reference to *all* its income prospects. Then the use of endowment funds would be the result of a comprehensive optimization. But there are some good reasons why university trustees might choose to stabilize their consumption from endowment alone.

One reason is that some major sources of other income are endogenous. Student fees, net of financial aid, are the principal example. Both fees per student and the number of students are matters of discretionary policy so that the trustees cannot regard student fees as exogenous and gauge their consumption plan to the prospects of income from this source. Rather they must regard as a principal criterion of endowment income the fixing of student fees and educational quality equitably across generations.

Gifts and grants for current use rather than for capital are the third major kind of income. There are two important reasons against capitalizing them. They are highly uncertain, and their fungibility is generally limited both in use and in time. It is reasonable, if imprecise, for the governing board to ignore "soft" money in calculating permanent income and thus to let total budgets fluctuate as soft money comes and goes.

I. A Formula for Endowment Income

I will now set forth a proposal for the definition of endowment income.[1] It is convenient to regard the university endowment as an open-ended mutual fund, with $S(t)$ the number of shares at the beginning of year t and $v(t)$ the value of a share at that time. At some arbitrary base period the number of shares $S(0)$ was equal to the market value and $v(0)$ was by definition equal to one. The number of shares changes if some are sold and the proceeds consumed, or if the university purchases shares in the market. Capital gifts can be conveniently regarded as buying into the mutual fund at the current share value, thus increasing the size of S. However, in the argument below I will take no account of such gifts.

---

* Yale University. The principal debt of this paper is to Donald Nichols, and it is so great that this should really be his paper, except that he would then nearly monopolize this session. His responsibility is too great to be denied, but at least I will absolve him of blame. The genesis of the approach I describe in this paper is an examination of Yale's definition of income undertaken by Nichols, Henry Wallich, and me in 1971. Our discussions led to the formula which Nichols proposed in the course of a general paper on the definition of income. Equation (14) of that paper essentially reappears as equation (1) below. He has recently prepared a revised version, "Measuring Income from Endowment." I would also express thanks to my colleagues Henry Wallich and William Nordhaus, to Richard Ferguson and Kenneth Codlin of the Yale University Treasurer's Office and to Ian Parker, who skillfully performed the calculations reported in Table 1.

[1] As stated in the introductory note, the major credit for the approach goes to Donald Nichols.

The basic formula for university income for year t is:

$$(1)\quad Y(t) = y(t)S(t) = \left\{d(t) + \left(\frac{g_t - i_t}{1 + g_t}\right)v(t)\right\}S(t)$$

Here

$d(t)$ is cash receipts of dividends, interest, and rents per share during year t, in current dollars—call it "cash yield" per share;

g_t is the estimate at time t of the expected rate of growth per annum of $d(t)$;

i_t is the estimate at time t of the expected rate of inflation of the cost of the academic activities supported by the endowment; and

$y(t)$ is permanent university income per share.

Unless it just happens that $g_t = i_t$, shares of the endowment will be sold, or conceivably bought, in order to meet the difference between calculated income Y and cash receipts dS. The change in the number of shares, ignoring gifts and assuming sales or purchases to be made at beginning-of-year values $v(t)$, is:

$$(2)\quad S(t+1) - S(t) = -\left(\frac{g_t - i_t}{1 + g_t}\right)S(t),$$

so that

$$(3)\quad S(t+1) = S(t)\left(1 - \frac{g_t - i_t}{1 + g_t}\right) = S(t)\frac{1 + i_t}{1 + g_t}.$$

If the estimate of growth g_t is correct, and if capitalization rates and expectations in security markets remain unchanged, both d and v will increase at the rate g_t. The ratio of cash yield per share to share value will remain constant. Under these circumstances,

$$(4)\quad y(t+1) = d(t)(1+g_t) + \left(\frac{g_t - i_t}{1 + g_t}\right)v(t)(1+g_t) = y(t)(1+g_t)$$

Using (3) and (4) we have:

$$(5)\quad Y(t+1) = y(t+1)S(t+1) = Y(t)(1+i_t)$$

That is, total dollar income will grow at the estimated inflation rate. If this estimate is correct, the university sustains its real consumption. Likewise, the dollar value of the endowment—still abstracting from gifts—will grow at the rate of inflation.

A. *Sensitivity of Formula to Securities Prices*

The formula is designed to permit a university to consume *recurrent* capital gains, but to avoid swings of income due to transient fluctuations in securities prices and changes in market discount rates.[2] How

---

[2] In addition to the smoothing implicit in (1), in practice it would also be desirable to smooth $y(t)$ in (1) in some way so that random components of the particular cash yield and share value of year t do not have excessive influence. Anyway, planning makes it necessary to set $y(t)$ before $d(t)$ and $v(t)$ are known. We can define $y^*(t)$ as permanent income and substitute y^* for y in formula (1).

There are several possible smoothing devices. Since any past y, projected forward at the growth rate, is an estimate of $y(t)$, one projection of $y(t)$ would be a sensitive is income, by this formula, to changes in securities values? Suppose that $v(t)$ falls by 1 percent while cash yield per share remains unchanged. Then income will fall by a fraction x of 1 percent, where x is the share of income obtained by realization of capital gains. Thus, if the university, following the formula, is relying on gains for one-fifth of its income, a fall of 25 percent in share value means a reduction of 5 percent in income.

If by coincidence $g_t = i_t$, the formula coincides with the old rule: consume cash receipts only. A university does not care how the market values the income streams it owns if it never intends to sell them. (This of course refers to the calculation of income, not to the tactics and strategy of portfolio management.)

If the academic inflation rate i exceeds the growth rate g, the university is a net purchaser of securities. In this situation, a rise in interest rates, making income streams cheaper to buy, is actually good news.

B. *Income in Relation to Fundamental Determinants*

The rationale of the formula can be seen from a somewhat different angle by considering the normal relationship of securities values to the internal rates of returns and saving propensities of business firms. Let s be the share of earnings reinvested, while $1-s$ is distributed as cash yield—dividends and interest. Let r be the growth rate of earnings e and cash yield d per "share." Then the growth rate, of which g_t is meant to be an estimate, is sr, and the value of a share is:

$$v = \frac{d}{\rho - sr} = \frac{e(1-s)}{\rho - sr}, \tag{8}$$

where ρ is the market discount rate.

From this it follows that

$$y = \frac{(\rho - i)e(1-s)}{\rho - sr}, \tag{9}$$

and if the market discount rate ρ is equal to the internal rate of return r,

$$y = \frac{(r-i)e}{r}. \tag{10}$$

Equation (10) says that under idealized conditions the university can consume the noninflationary fraction of the earnings of the businesses whose securities it holds. The remainder, $(i/r)e$, must be plowed back to enlarge the endowment enough to keep up with inflation. It may be that the businesses themselves do this reinvestment, that $s \geq i/r$. If so, the university doesn't need to save on its own and may have to realize part of its income by selling securities. But if $s < i/r$, then the university will have to save part of its nominal cash income.

Equation (10) also indicates that if $\rho = r$ the definition of income is independent of the saving rate s. Securities values depend upon earnings, not on how they are divided. Although cash yield d declines when the saving rate increases, this is offset in (1) by the increase in the growth rate.

The business saving rate is not neutral when ρ and r differ, as in (9). If $r > \rho$ then greater s will make y bigger. The reason is that a larger value of s means a higher growth rate and more future profits for the market to overvalue. (If $r < \rho$ the opposite is true.)

II. Estimation of Growth Rate of Earnings

One approach to implementing formula

---

weighted average of such individual estimates:

$$y^*(t) = \sum_{u=t-n}^{t-1} y(u)(1+g_t)^{t-u} w(t-u) \tag{6}$$

$$\sum_{u=t-n}^{t-1} w(t-u) = 1$$

Another possibility is

$$y^*(t) = \alpha y^*(t-1) + (1-\alpha) y(t-1) \tag{7}$$

(1), modified by smoothing, would be to guess the parameters each year, using current forecasts and analyses of the economic climate and the securities markets. The formula then has at least the merit of organizing such guesses, making explicit the parameters on which judgment is required and their influences on the final result.

Alternatively, parameter estimates may be sought in the university's own experience, or indirectly by first estimating economy-wide parameters and then correcting them for the particular circumstances of the institution. So far as g_t, growth rate of cash yield, is concerned, the difference arises from the fact that historically the university, for better or worse, has not owned a cross section of the securities of American enterprises. Moreover, the composition of the portfolio as between bonds, low-pay-out stocks, and high-pay-out stocks may have fluctuated widely. Therefore the actual trend of the university's cash yield reflects portfolio tactics rather than the growth that could be expected from a portfolio of stable composition. It would be possible, but potentially misleading, to estimate the historical g_t for the university's current portfolio, on the contrary-to-fact assumption that it had been held over a long period in the past.

The alternative is to estimate a real growth rate for the economy as a whole, G_t, and then to correct it for the deviation of the university's portfolio from an economy-wide cross-section portfolio. Imagine, for example, a reference portfolio which year after year is a claim to a fraction of the interest and dividends actually paid out by American corporations. We are looking for the trend of cash yield *per "share"* of the reference portfolio, so we must not count any of the national growth of interest and dividends which reflects net issues of securities by American corporations, old and new.

We want to know how cash receipts would have grown for an investor with a constant number of "shares," an investor who did not receive additional capital gifts enabling him to buy new securities. His share in the American economy dwindles, but he reshuffles the composition of his portfolio so that he always owns a cross section of the claims outstanding. For example, the reference investor increased his relative holdings of bonds as corporations shifted from equity to bond finance in recent years.

Since the trend of real cash yield per share is more likely to be stable than the trend of nominal cash yield, the parameter estimated from time series would be $G_t - I_t$, the economy-wide counterpart of $g_t - i_t$. This could be done by regression of deflated cash yield per share against time, and as new observations become available with the passage of time the estimate of $G_t - I_t$ would be regularly revised. To free the trend estimate from cyclical influences on real cash yield, the regression should also include a cyclical indicator like the unemployment rate or the Okun *GNP* gap. Calculations of this kind for a reference portfolio are reported in Table 1.

Given an estimate of $G_t - I_t$, how can we move to $g_t - I_t$, the specific real growth rate for the university endowment? (The remaining correction for $i_t - I_t$, the excess of the academic rate of inflation over the general rate of inflation, does not seem so difficult.) One procedure would be to assume that the real internal rate of return is the same for the university portfolio of enterprises as for the reference portfolio and to attribute the whole difference between g and G to differences in the earnings retention rates of the two sets of companies:

$$\frac{G_t - I_t}{S_t} = \frac{g_t - I_t}{s_t} \tag{11}$$

TABLE 1—ILLUSTRATIVE CALCULATION OF PERMANENT INCOME ON REFERENCE PORTFOLIOS

Year	Aggregate Dividends & Interest[a] (billions of $) (1)	Cash Yield per Share[b] (billions of $) (2)	Real Cash Yield per "Share"[c] (billions 1948 $) (3)	Permanent Cash Yield per "Share"[d] (billions 1948 $) (4)	Estimated Growth Rate of (3)[e] G_t-I_t (5)	Market Value per share[f] (billions of $) (6)	Number of Shares in Portfolio[g] After Purchase or Sale (7)	GNP Price Deflator[h] (1948=1.0) (8)	Permanent Real Income (billions 1948 $) (9)
1948	7.85	7.85	7.85	7.55	−.004	153.9	1.004	1.000	6.96
1949	8.18	7.82	7.87	8.21	−.004	135.7	1.008	0.994	7.72
1950	9.70	8.90	8.83	9.20	−.004	151.1	1.012	1.008	8.70
1951	9.65	8.53	7.93	7.60	−.004	156.0	1.016	1.075	7.13
1952	9.83	8.32	7.57	7.02	−.004	161.1	1.020	1.099	6.56
1953	10.18	8.22	7.41	6.82	−.004	158.7	1.024	1.109	6.40
1954	10.88	8.42	7.48	7.61	−.004	186.8	1.028	1.126	7.15
1955	12.10	9.00	7.88	7.90	−.004	232.7	1.033	1.142	7.31
1956	13.05	9.40	7.96	7.79	−.004	241.5	1.037	1.181	7.23
1957	14.13	9.86	8.05	7.91	−.004	241.9	1.041	1.225	7.42
1958	14.30	9.80	7.80	8.46	−.004	257.1	1.045	1.256	7.99
1959	15.33	10.38	8.13	8.61	−.004	318.1	1.049	1.278	7.99
1960	16.45	11.04	8.51	8.89	−.004	314.6	1.053	1.298	8.86
1961	17.38	11.55	8.79	9.62	−.002	372.8	1.056	1.314	9.86
1962	19.33	12.72	9.57	10.13	.001	369.1	1.057	1.329	10.99
1963	21.00	13.80	10.25	10.79	.001	418.7	1.056	1.347	13.36
1964	22.88	14.97	10.94	11.33	.006	470.9	1.049	1.368	15.86
1965	25.83	16.84	12.08	12.11	.011	524.8	1.038	1.393	19.61
1966	28.15	18.28	12.76	12,36	.018	506.2	1.020	1.432	20.89
1967	30.33	19.59	13.26	12.75	.023	556.6	.997	1.477	22.84
1968	33.80	21.68	14.11	13.46	.026	622.8	.970	1.536	24.47
1969	37.30	23.78	14.78	14.02	.029	635.9	.943	1.609	24.40
1970	39.25	24.68	14.56	14.56	.029	565.1	.915	1.695	23.10

[a] Dividend and bond interest payments of *U.S.* private corporations.

[b] The claim to (a) represented by securities outstanding in 1948 or by securities replacing retirements. Excludes claims represented by net new issues of securities. For bonds, this calculation required some simple arbitrary assumptions concerning the distribution of maturities and coupons in the initial stock and subsequent issues.

[c] =(b)/(h).

[d] (c) adjusted for cyclical fluctuation to approximately 4.5 percent unemployment.

[e] Estimates 1948–60 from regression 1948–59. Thereafter, estimate from year t from regression 1948 −(t−1).

[f] Market value of the claim described in (b).

[g] The cumulative result of following the formula for permanent income.

[h] =(i)+(e) (f) (g)/(h).

After computing the two retention rates, S_t and s_t, we can estimate g_t from (11).

As Donald Nichols has forcefully pointed out to me, however, the assumption that "growth" companies are simply high-retention companies is empirically wide of the mark. Since the market does not equalize real rates of return across companies, growth stocks are often simply those that have high internal rates of return, regardless of pay-out ratios.

An alternative, suggested by Nichols, is to assume that the market knows what it is doing, so that differences in growth rates can be discerned in differences of cash yields relative to share values. If share value is $v=d/(\rho-g)$ for the university portfolio and $V=D/(\rho-G)$ for the reference portfolio, then

$$g-G=\frac{D}{V}-\frac{d}{v}. \tag{12}$$

III. Portfolio Management and the Income Definition

Neither of these procedures allows for any special skill on the part of the portfolio manager. He is assumed to do no better than the reference portfolio, though his tastes in securities may be different. If the portfolio manager is worth his salt and his investment performance is systematically and recurrently superior, it should be reflected in current as well as in future income. This could be done by monitoring the growth rate of $v(t)$, market value per share, and by including some fraction of its excess over the inflation rate in the annual estimate of g_t.

A modified version of the old rule could be defended as a low-risk portfolio strategy. This would be to put enough of the endowment into bonds and high-dividend stocks so that $g_t = i_t$. Then no securities need to be bought or sold to meet the formula definition of income, and the university is insulated from fluctuations of capital values. A university is interested in the security of its income, not in the stability of the endowment's market value. In choosing investments, portfolio managers should be willing to take considerable risk of capital value for an assured and growing cash yield.

REFERENCE

D. Nichols, "A New Measure of Income from Wealth," Cowles Foundation Disc. paper no. 321, Nov. 1, 1971.

A Plan for the Definition of Endowment Income

By JAMES M. LITVACK, BURTON G. MALKIEL, AND RICHARD E. QUANDT*

The conventional financial rule under which institutions deriving at least part of their income from securities have traditionally operated is that dividends and interest may be spent but not capital gains. In recent years the conventional wisdom has been attacked, partly on the grounds that it ill serves the institutions in question.[1]

The purpose of this paper is (a) briefly to examine the principal reasons for employing a definition of endowment income different from that implied by the conventional spending rule, (b) to discuss the implementation of the new definition which was designed at Princeton University and contrast it with some alternatives, and (c) to indicate the new spending rule's operation with the aid of some sampling experiments.

I. The Definition of Endowment Income

The definition of endowment income should be consistent with the following criteria: (1) the definition should seek to make investment management independent of the spending decisions of the university; (2) it should imply a spending rule that protects the real value of the corpus of the endowment fund; and (3) it should seek to make spendable endowment income for the university reasonably stable from year to year. We shall show that the conventional definition fails to meet the first two requirements.

Independence of Spending and Investment Management. The objective of investment management should be the maximization of *total* rate of return, defined to include changes in capital values as well as interest and dividend receipts, subject to the constraint that the risk level assumed is not inconsistent with the need to preserve the corpus of the fund. It is the size of the total return, not its composition, that matters if the university is to achieve the largest stream of resources over time.[2]

In institutions using the conventional definition of endowment income—and therefore prevented from spending capital gains—concern for maximizing spendable income might induce the investment manager to increase the proportion of the portfolio invested in fixed income securities, whether or not this policy was consistent with the objective of maximizing total returns. Thus, traditional spending rules may easily force the university into an inappropriate portfolio-management policy.

Suppose, however, that investment policy was set according to wealth-maximizing objectives, and that this approach led a university to put the major portion of its endowment funds into stocks that pay low dividends but are expected to produce sizable capital gains.[3] Since traditional rules dictate that none of the capital gains can be applied to current needs, the university is forced to allocate a large share of the total return on endowment to future needs.

---

* Financial Research Center, Princeton University. We are indebted to Mark R. Killingsworth for helpful comments.

[1] For example, such institutions include Amherst College, Princeton University, the University of Pennsylvania, and the University of Rochester. The American Economic Association abandoned the conventional definition some time ago (see Reports of Committee on Financial and Investment Policies).

[2] Robert M. Haig and Henry C. Simons presented the classic argument that all accretions to an institution's wealth enlarge its capacity to command goods and services.

The word "forced" is used advisedly, since the all-important allocation decision as between present and future needs is made by an automatic mechanism that allows no room for conscious decision. This is not to argue that the traditional definition of spendable income inevitably assigns too high a proportion of total endowment return to future needs. Indeed, during the decade of the 1960's, when new money flowed abundantly to universities, it was possible to meet the most pressing demands of the moment while allowing capital gains to be added to the corpus of the endowment fund. That happy situation no longer prevails for universities, and somewhat lower savings rates may well be more appropriate and, indeed, necessary for the 1970's.

In sum, the major objection to the traditional spending rule is that it gives rise to an insoluble dilemma: If the endowment is invested for maximum total return, only by accident will the amount of dividends and interest earned in a given year be consistent with the amount that ought to be spent on the basis of general policy considerations. But if a conscious decision is made on the amount of endowment income that ought to be available for spending, and if this decision is allowed to influence investment policy, only by accident will the composition of the portfolio be consistent with the objective of maximizing total return. Any definition of income should seek to separate questions of resource allocation as between present and future from the conduct of investment policy.

Preservation of the Real Value of the Endowment. University trustees have a clear obligation to protect the corpus of an endowment fund, and this obligation is often given as the justification for a policy of prohibiting the spending of capital gains. But the traditional definition of income implies that we have different standards for the preservation of the capital for stocks and bonds.

For bonds the implicit standard is that we should preserve the dollar value of the principal amount. For stocks the rule implies that we preserve the same number of shares or, if there are stock splits and dividends, that we preserve the endowment's proportional equity in the companies owned. Thus, the traditional rule may not effectively preserve principal at all. Suppose that the university held a portfolio consisting entirely of bonds. Under the traditional rule, the university can spend all of the interest income from the bonds, and capital is "preserved" so long as the principal amount of the bonds is kept intact. But, if there is any increase in the price level, the real value of the portfolio declines and the purchasing power of the endowment has not been preserved.

On the other hand, suppose the institution holds a portfolio of non-dividend-paying stocks. It may well be that the overall return (including capital gains) for this institution is far higher than for the university that holds its portfolio entirely in bonds. Nevertheless, the second institution is defined as having no spendable income whatsoever and all capital appreciation (overall return) is considered to be part of the corpus of the fund. Here capital is preserved with a vengeance. As long as the overall return from the stock portfolio is greater than the rate of inflation, the institution will be adding to the real value of the endowment, not merely preserving it.

The point is that treating only dividends and interest as spendable income is an unreliable method of preserving the corpus of the endowment fund. The traditional rule may either undercompensate or overcompensate for changes in the relevant price level. What is needed is a more direct and reliable method of preserving the real value of the endowment.

---

[3] M. H. Miller and Franco Modigliani have demonstrated that in perfect capital markets a company that retains $1 of earnings will produce $1 of capital gains for its stockholders. For investors such as universities that pay no taxes, the latter is just as good as the former.

Stability of Spendable Income. Spendable income for the university should be relatively stable from year to year. In view of the fixed nature of so many of the financial commitments made by any university, it is especially important to minimize the likelihood of sudden reductions in income.

The traditional spending rule is consistent with our third criterion, since dividends and interest receipts are reasonably stable from year to year, provided there is no change in the composition of the portfolio. The university can estimate with a high degree of accuracy what spendable income will be from year to year. It is well known, however, that changes in capital values do not occur smoothly over time. Thus, if an alternative definition of income is employed, it is important to assure that spendable income from year to year is reasonably predictable and stable.

Toward an Alternative Definition of Income. We have implicitly argued above that dividends and interest plus capital gains is the proper definition of income. More specifically, we define spendable income as $Y = [(d+c)-g]M$, where d is the long-run rate of dividends (interest), c is the long-run rate of capital gains, g is the desired rate of growth of M, and M is the market value of the portfolio. L. Fisher and J. H. Lorie have estimated that, for long periods from the 1920's to the 1960's, the long-run total rate of return from stocks was a little over 9 percent, which we took to be a realistic estimate of $(d+c)$.[4] We then assumed that prices would rise by 5 percent annually, so that the university would desire to increase the nominal value of its endowment by 5 percent annually to keep constant its real value (i.e., $g=0.05$). Thus, spendable income is 4 percent of M, as the specific plan presented below assumes.

II. The Princeton Plan and Some Alternatives

The plan proposed for Princeton makes use of a 9 percent expected total return (including dividends and capital changes) to generate a spendable income rule. To insure the body of the endowment against inflation, a 5 percent rate of anticipated inflation is deducted from the total return to arrive at a spending rate. The 4 percent spending rule is applied to a moving average of past endowment values to assure a reasonable degree of stability in spending. Moreover, the plan includes a "tracking rule" that reduces the rate of spending should total return remain below the long-term trend for a long period of time.

Essentially, Princeton computes a capital value which is a weighted average of the past six semi-annual values of endowment.[5] The weighting formula increases all past values at a compound annual interest rate of 5 percent; i.e., if the endowment value six months previous in period t−1/2 is 100, the number entering the average is 100 times 1.025. The "rate of return" is defined as the average of the past six semi-annual rates of (total) return. The product of the capital value of the endowment and the rate of return represents the income to be allocated. Five percent of the capital value is allocated to endowment to preserve the real value of the fund, 4 percent of the capital value is allocated to "spendable income," and the remainder is added to (or subtracted from) a "stabilization fund."

---

[4] Donald Nichols argues that dividends and interest plus capital gains is an invalid measure of spendable income and that Y should be defined as $[d+(rs-g)]M$, where r is the rate of return on corporate investment and s is the corporation's retention rate. But if $g=0$, then rs is simply the (long-term) rate of growth in the portfolio's market value, and the problem is to find the best empirical approximation to rs. Capital gains in any short-time period are an unsatisfactory approximation (which is why we use long-term values), and retained earnings are even less satisfactory: accounting earnings are ill defined, and, as William J. Baumol, Peggy Heim, Malkiel, and Quandt have shown, retained earnings seem to bear little relationship to the estimated growth of profits.

[5] A fuller description of the plan is contained in the Faculty-Administration Committee Report.

More precisely, the plan allows 4 percent spending when the stabilization fund exceeds 50 percent of its "full value," where "full value" is defined as the sum of the three preceding years' "spendable income." When the stabilization fund is below 50 percent of "full value," spending is scaled down by schedule to 3.2 percent. This base figure of 3.2 percent, which has been the rate of dividend and interest income, becomes operative if the stabilization fund is less than 5 percent of full value. It should be noted that this stabilization fund is simply an accounting tool that allows Princeton to make sure the earnings rate, long term, is not less than 9 percent if the university is to spend 4 percent.

Princeton's long-term investment record has averaged close to a 10 percent rate of return. Since dividends and interest have averaged only 3.2 to 3.4 percent, our plan would allow a higher rate of spending than traditional rules. The use of averages prevents sharp, short-term market fluctuations from too greatly altering spending. The 5 percent inflation factor seems to offer a reasonable degree of protection against declines in the real value of the endowment in the long run.[6] Finally, should our long-term assumption of a 9 percent total return prove too optimistic, the plan automatically reduces spending to a rate comparable to the traditional dividend plus interest rule.

Alternative plans for the redefinition of endowment income resemble the Princeton plan in some respects, although the latter has several novel features. (1) It explicitly contains an allowance for estimated inflation in the definition of spendable income —an allowance which can be adjusted in the light of future experience. (2) It explicitly scales down spendable income if total return falls short of its anticipated level. (3) It makes no distinction between realized and unrealized gains and thus is consistent with a portfolio strategy of low turnover. (4) It avoids the measurement difficulties inherent in the use of retained earnings in defining endowment income. (5) It is applied to all funds, not just "funds functioning as endowment."[7]

III. Operation of the Spending Rule

Table 1 shows how the Princeton plan would have operated over the 1959–1971 period, given Princeton's actual investment experience. Had the plan been in operation, Princeton would have been able to spend somewhat more income than was possible under the traditional spending rule. This is because Princeton's investment policy has been to weight the portfolio in favor of common stocks with good growth prospects but with low rates of dividend return. The proposed plan would thus reduce pressures on the portfolio managers to change the composition of the portfolio toward bonds, which have a much higher current rate of return. Despite the somewhat larger spending rate in

---

[6] The relevant price index is, of course, that of the market basket of goods and services purchased by the university, which may well be different from standard price indices. Whether the appropriate index will rise faster than some standard index is not clear. There is some reason to think it will go up less fast since the salary component of the index is at least under the partial control of the university. On the other hand, the activities of universities may well be those least likely to achieve productivity gains.

[7] "Funds functioning as endowment" are those that are not legally restricted as to the spendability of the corpus of the fund. Applying different spending rules to different portions of the whole endowment will change their relative proportions in the long run.

TABLE 1—SPENDABLE INCOME, GIVEN ACTUAL INVESTMENT EXPERIENCE, FOR PRINCETON UNIVERSITY

(Millions of Dollars)

	Actual Dividends and Interest	Plan Results
1956–57	4.3	—
1957–58	4.9	—
1959–60	5.3	—
Plan begins		
1959–60	5.7	5.2
1960–61	6.2	6.1
1961–62	6.6	7.2
1962–63	7.2	8.2
1963–64	7.7	9.0
1964–65	8.5	10.0
1965–66	9.6	11.1
1966–67	10.9	12.4
1967–68	11.7	13.4
1968–69	12.7	14.3
1969–70	14.0	15.6
1970–71	15.0	15.9
Endowment at end of fiscal year 1970–71	440.5	426.1

Note: Simulation of plan assumes stabilization fund is initially set at 25 percent so spending rate is 3.5 percent initially.

most years, the corpus of the endowment is protected against a 5 percent inflation. Moreover, our averaging technique insures that spendable income from year to year is relatively stable. Indeed, spendable income shows a steady growth and never declines over the sample period.

We also performed a number of simulation experiments to see how the plan might operate in the future. We assumed both higher and lower total rates of return than had been experienced in the past and also more and less variance in stock prices than over the past fifteen years. We generated fictitious histories of prices with the indicated mean and variance. Under a wide variety of circumstances the proposed plan is consistent with the *desiderata* set up in the start of the paper; it allows slightly increased spending and produces a reasonably smooth spending stream.[8] In sum, it would appear that the Princeton plan is a serviceable one that can make an important contribution to the more effective use of the University's resources.

REFERENCES

W. J. Baumol, P. Heim, B. G. Malkiel, and R. E. Quandt, "Earnings Retention, New Capital, and the Growth of the Firm," *Rev. of Econ. and Statist.*, Nov. 1970, *52*, 345–355.

L. Fisher and J. H. Lorie, "Rates of Return on Investments in Common Stocks," *J. of Bus.*, Jan. 1964.

R. M. Haig, "The Concept of Income—Economic and Legal Aspects," in R. M. Haig, ed., *The Federal Income Tax*, New York 1921.

M. H. Miller and F. Modigliani, "Dividend Policy, Growth and the Valuation of Shares," *J. of Bus.*, Oct. 1961, *34*, 411–433.

D. A. Nichols, "A Measure of Income for Perpetual Institutions," Univ. of Wisconsin, *SSRI* workshop series no. 7214, July 1972.

H. C. Simons, *Personal Income Taxation*, Chicago 1938.

Faculty-Administration Report, *The Definition of Endowment Income*, Princeton, Feb. 1970.

Reports of Committee on Financial and Investment Policies of the American Economic Association, *Amer. Econ. Rev., Proc.*, May 1967, *57*, 711–714, and May 1969, *59*, 586.

---

[8] It should be noted, however, that the larger the assumed variance of stock prices, the less the final value of the endowment. This is so because of a phenomenon analogous to reverse dollar-cost averaging which means that slight adjustments in spending rates will be necessary, depending upon the expected variance of stock prices.

DISCUSSION

ROBERT EISNER, Northwestern University: The request that I concentrate my remarks on "The Investment Income Formula of the American Economic Association" leads me to recall early years of the Sino-Soviet dispute. Chinese attacks then were directed against the revisionist clique in Yugoslavia while Russian propaganda blasted Albania. My topic raises some important policy issues and some subtle but significant points of theory which apply most importantly neither to little Albania nor to the miniscule portfolio of the American Economic Association (*AEA*), but rather to the great bulk of private colleges and universities in the United States, and a number of the public institutions as well. In a broader sense we come to grips with the basic concepts of income and wealth, welfare, and intergenerational allocation and optimalization.

The current problem of endowment income of educational institutions had its major and modern genesis in the 1950's. Immediate needs in terms of faculty salaries and educational services generally seemed enormous. Yet trustees and some administrations of private universities revealed a set of preferences which appeared to some of us most appropriate for managers of an investment trust. In many cases they had acquired portfolios including substantial, if insufficient, numbers of low payout "growth" stocks. But those persons accounting for university resources and significantly influencing the expenditures were willing to count as income to be spent for current educational purposes only dividends and interest and ignored capital gains. A university, it seemed, could improve by leaving the next generation a higher endowment than the amount it received, not by spending to leave a higher quality faculty, a greater investment in research, or whatever else a college and university was supposed to be about.

Some enlightened university administrators, such as W. Allen Wallis, the distinguished chairman of this session, were properly concerned. It was thought that the *AEA*, with its own meager portfolio, might set an appropriate example for the colleges and universities for whom endowments were a major item of wealth and income. An *AEA* Committee on Financial Investment Policies, under the impeccable chairmanship of Milton Friedman (and including the present author) recommended and secured adoption by the Association of a formula which made its "income from equities in each year equal to dividends received in that year plus one-third of market appreciation of that year (adjusted for price-level change) plus one-third of adjusted appreciation in the preceding year, plus one-third of adjusted appreciation in the second preceding year" (see *Amer. Econ. Rev., Proc.*, May 1969, *2*, 586).

Now, as Donald Nichols points out, while there is merit in including some portion of real capital gains in income, it is not clear that the *AEA* formula, which counts in current income a simple average of real capital gains on equity over a three-year period and real capital gains of the current period on bonds, is universally appropriate. Following Sir John Hicks, if one defines income as what one can consume during a period and still be as well off at the end of the period as at the beginning, and if "being as well off" relates to expecting at the end of the week to have as much future "welfare" as was expected at the beginning of the period, the amount of capital gains which can be consumed without reducing expected utility in the future depends upon an intertemporal utility function. If the intertemporal utility function is maximized by consuming indefinitely at a constant rate, then no part of capital gains due to reductions in the rate of discount of expected future receipts can be included in income. For example, if a portfolio consists of a million consols each paying one dollar per year to the holder forever a change in the market rate of interest or discount from, say, 10 percent to 8 percent, thus raising the value of the portfolio from $10 million to $12.5 million has no effect whatsoever on the future stream of earnings, which remains at the initial one million dollars annually.

But the specific formulation by Nichols leads us, and James Tobin moves explicitly, to undervalue the present and, hence, condone or encourage the old investment trust concept of handling a university portfolio. For Nichols suggests a definition of income equal to optimal consumption such that $C_0=R_0+(G_R-G_C)V$, where V is the value of the portfolio, G_R and G_C are respective growth rates of receipts and consumption, and consumption plans are presumed to "involve exponential growth, though perhaps at a low or even zero rate." Tobin is specific, declaring: "Trustees of an endowed university . . . assume the institution to be immortal. . . . Sustainable consumption is their conception of permanent endowment income. . . . In formal terms, the trustees are supposed to have a zero subjective rate of time preference." Hence, via Nicholas, Tobias arrives at his basic formula (1), $Y=[d+(g-i)V]S$, which makes endowment income (Y) equal to current cash receipts of dividends, interest, and rents per share (d) plus the product of the value of a share (V) and the difference between the expected growth per annum of cash receipts per share (g) and the expected rate of inflation (i), all, of course, multiplied by the number of shares (S).

Calculating or estimating the expected rate of growth of cash receipts per share then becomes critical. Tobin does this by taking *U.S.* securities as a whole as a representative portfolio and regressing deflated cash yield per share against time. In his estimates he comes up with rates of growth which are actually (very slightly) negative to 1961 and finally reach a maximum of +2.9 percent in 1969 and 1970. Thus, the Tobin formula has counted in income none of the enormous capital gains in American securities in the decade and a half after World War II and relatively little since.

Yet Tobin's calculations show an increase in the market value per share from $153.9 billion in 1948 to $565.1 billion in 1970 which, after application of the *GNP* price deflator he uses (and despite the sharp drop in market value per share from 1968 to 1970), still involves an increase of 116 percent in real value of the "reference portfolio," over 3.5 percent per year. Since the cash yield, which constituted virtually all of Tobin's income, averaged only about 3.5 percent during 1948–1970, his measure was little more than half of total income including real capital gains.

The impact of the Princeton formula, as reported by James Litvack, Burton G. Malkiel, and Richard E. Quandt, is similar, but here the conservatism is explicit. Endowment income (spendable income) is set at a maximum of 4 percent of the market value of the portfolio, with a sliding scale that reduces the ratio to as low as 3.2 percent if growth is below expectations. In my own university, with guidance from an enlightened president who is both a distinguished economist and student of certain problems of myopia, without any change in accounting procedure we have moved towards appropriating to income enough of capital gains to yield a 5 percent return on portfolio. This would appear slightly more liberal than the Princeton formula and, of course, decidedly more liberal than Tobin's equation would have been if applied up to 1970.

But what can we make of all of this in terms of the actual numbers or the theory that has led us to them? Are true real rates of return only 3 and 4 percent or even 5 percent? As Litvack, Malkiel, and Quandt remind us, Fisher and Lorie have estimated long-run total rates of return from stocks as a little over 9 percent, and this over a period when the inflation rate subtraction could hardly have been much more than 2 percent. Indeed, I can hardly imagine a portfolio manager of a bank, a mutual fund, or a university who would admit that 3 or 4 percent was all the return he could expect. But if the true rates of return are more than this, application of the formulas of any of the three papers under discussion would reinforce and perpetuate the university practice of using endowments to build forever for the future. Jam tomorrow, but never jam today!

There are two directions in which we appear to be going astray. One relates, in the Tobin formulation at least, to the concentration on cash payouts rather than total earnings. The other involves the fundamental issue of time preference.

One reason, I submit, why Tobin picks up so little of capital gains is that he perceives capital growing with the growth in cash payout, d, which is quite reasonable on his assumption that a constant proportion of earnings $(1-s)$ is actually paid out. But high marginal rates of taxation for most recipients of dividends and interest and high corporate profits tax rates, along with increased opportunities to lower reported corporate earnings by larger charges for depreciation and other accounting costs, have led to several distorting elements in the ratio of cash yield to true total business earnings. While the growth in net earnings has seemed to slow with accelerated depreciation charges, it is quite possible that, on an accounting basis consistent over time, earnings have grown more rapidly than dividends. For the period when Tobin found a negative rate of growth of real cash payout, for example, real earnings may have been growing at a substantial rate while the ratio of earnings paid out was declining at an equal rate. As a consequence, earnings paid out remained about constant (or actually declined) even while, from 1948 to 1961, the real market value per share rose at almost 5 percent per year.

We might, then, work with the growth in earnings rather than the growth in cash payout, but this would appear likely to raise even more difficulty. For the cash payout is at least an observable market phenomenon. Earnings figures, on the contrary, are bedeviled by accounting vagaries distorted by tax considerations. We would do better to rely on market figures, but these should be the directly relevant ones, the capital gains (or losses) reflected in market evaluations for securities in the portfolio. And that is indeed what the *AEA* formula does.

But there is still another argument for accounting for income in the manner of the *AEA* rather than along the lines of the Nichols-Tobin formulation. This relates to that basic matter of time preference.

Why should we act as if a university or any other institution is permanent? We know that nothing is permanent. We know that we are all mortal. We know that needs change and we can anticipate that the needs for endowment income can change. We know that populations move and that particular universities may be in less demand in the future (while others may or may not be in greater demand). Indeed, almost nowhere in human behavior or in economic activity do we show a zero rate of time preference. We are always giving more weight to the present than to the future. Because institutions are mortal, or because we may expect to have more tuition, government support, endowment income, or gifts in the future, the probability of requiring income from our present endowment is likely to decline over time. We should, therefore, not expect to maximize expected utility by providing for equal expenditures from endowment over all years in the future. Formally, we might write

$$P(\mathrm{t}+1)\frac{\partial U}{\partial C(\mathrm{t}+1)} = P(\mathrm{t})\frac{\partial U}{\partial C(\mathrm{t})},$$

where P is the probability of needing the consumption expenditures, C, in the indicated years t and t+1. We thus equate *probability-weighted* marginal utilities of consumption (or educational expenditures from endowment income) or equate marginal rates of substitution to ratios of probabilities.

But if this is so, and if these probabilities are a declining function of time, or if we have positive time preference for any other reason, then we may reject the argument cited earlier that we should not spend capital gains resulting from declines in the rate of discount of an unchanged expected stream of future earnings. If we are not planning to use these capital gains forever at an even (rather than declining) rate, then we can begin to spend some of them. And that is precisely what would flow out of Friedman's function relating consumption to permanent income and the rough pragmatic application of this in the *AEA*'s measure of its income.

If we are to maximize subject to our preferences, we may well ask whose preferences and what they are. Current students and faculty may want a measure that provides more income and expenditures now, and ambitious administrators, who measure their success by the progress of their institutions rather than its endowment, may reason similarly. But, it is argued, trustees are responsible for the immortal soul of the institution, and they and, most importantly, donors are interested in providing something immortal. That, I submit, is not clearly true. Many donors, for example, give money for buildings: a new medical center, a gymnasium, a business education complex. But how many times has a donor made the following statement? "Here is $10 million for a new building to be named after me, but I do not want you to spend all of it on the building now; rather give me an $8 million building in my name and set aside $2 million which will accumulate at compound interest so that you can always have a fund to replace the original building and its successors, as they wear out, thus retaining my name for posterity." That has not happened to my knowledge at Northwestern University nor generally, I suspect, even been thought about. Donors, I submit, like almost everyone else in our society, while aiming for a bit of immortality, do not expect anything to last forever and act accordingly.

The principles laid down by Nichols and Tobin have much to commend them and it is certainly true that there is no unique, best, universal concept of income. But there is good reason to work with a concept of income which recognizes gains in the present as they occur, which means as the market recognizes them, and which guides us to provide adequately for the future without short-changing present needs. We clearly do not want to squander all of our capital in expenditures of this week or year. But neither do most of us have any good reason to reserve our capital to provide equal expenditures forever and ever and ever. If the *AEA*, like Albania, exists in a thousand years, we should anticipate that the fine economists of that millennium, as well as those who have built up the endowment of the Association in the immediately preceding period, would provide amply for its needs. And the same argument should apply for Harvard, Yale, Princeton, Wisconsin, Northwestern and Slippery Rock.

STANLEY W. BLACK, Vanderbilt University: These papers contain a substantial amount of guidance and wisdom for university treasurers and boards of trustees. The numerical embodiment of that wisdom provided by each of these papers raises a number of questions, however.

Let me recapitulate Nichols' basic argument once more. Capital gains can properly be counted as income for a perpetual institution only to the extent that they reflect profits from reinvested earnings of corporations. Other capital gains, such as those due to general inflation, changes in the expectation of future profitability of firms, or changes in the market's rate of discount for future profits, should not be counted as real income, since their effects on an institution's future consumption stream depend on the shape of that consumption stream and in any case are nonrecurrent.

So the question is: What portion of capital gains may properly be consumed without impairing the real value of the institution's endowment? The other two papers approach this question by means of Nichols' formula, but each supplies the parameters in a different way. Nichols' formula can be written

$$C = D + (rs - g)MV = (d + rs - g)MV,$$

where C is spendable income, D is dividends and interest, MV is the market value of the portfolio, d is the dividend yield, r is the internal rate of return of corporations, s is the retention rate, and g the desired growth rate of nominal spendable income and the market value of the portfolio (i.e., the rate of inflation).

Princeton assumes that the total return $d+rs=9$ percent, as estimated by Fisher and Lorie's study of stock prices. The inflation rate g is taken to be 5 percent, presumably based on current experience, yielding a spending rule of 4 percent of market value. This rule is smoothed by applying it to a five-year moving average of market values and, most crucially, is hedged by providing a stabilization fund. If actual total return exceeds 9 percent and the stabilization fund fills up, spending can (or should) exceed 4 percent. On the other hand, if actual total return is less than 9 percent, the stabilization fund will eventually run dry, and spending will return to the faith of our fathers: "Don't touch capital."

Tobin approaches the estimation problem differently. He estimates *rs* from its definition: the trend rate of growth of cash yield per share. As Nichols has shown, this will be the same in the long run as the rate of growth of share prices if and only if the internal rate of return r is the same as the market discount rate i. Tobin adds the *current* dividend yield to this estimate of *rs* to get the total rate of return $d+rs$. Then he deducts the *trend* rate of inflation from this total return to get his measure of income per share.

Using subscripts c for current and t for trend, one can thus write the two sets of parameters as:

$$\text{Princeton:} \quad (d + rs)_t - g_c$$

$$\text{Tobin:} \quad d_c + (rs)_t - g_t$$

This summary suggests the following comments:

(1) Since the inflation rate is not one of the "great constants" of economics and in fact can be whatever the monetary authorities choose, the current inflation rate, or some judgmental combination of recent inflation rates, will probably be a better figure than the historical trend used by Tobin.

(2) The Princeton measure of total return is conceptually different from the dividend yield plus rate of growth of dividend yield used by Tobin and Nichols. The Princeton measure in fact includes all of capital gains, albeit averaged over a long history since the Fisher-Lorie study covered a long period. Thus *most* of the capital gains included in the 9 percent estimate must have been due to retained earnings and inflation, not changes in expected profits rates or market discount rates.

(3) The specific period over which Tobin computes his estimates of *rs*, the rate of growth of cash yield per share, *may* be dominated by a change in market discount rates. The striking thing about Tobin's estimates of "permanent endowment income" is that they are so often *below* actual dividends plus interest. In fact, Tobin's estimate is below the conventional figure from 1948 to 1964, rising above dividend and interest only from 1965 to 1969. His message seems to be that universities were too prodigal in the 1950's and early 1960's. An important factor in his estimates is the decline in *real* dividends per share from 1950 to 1961, which imply that $rs-g$ is *negative* over that period. It is doubtful that inflation is the culprit in this case. Rather, it appears to be some change in parameters over this period that resulted in falling real dividends per share. Tobin's own figures show that the dividend yield declined from a peak of 5.89 percent in 1950 to a low of 3.10 percent in 1961. This matches Standard and Poor's figures and mirrors a reverse move in stock prices. It seems likely that the 1948–61 period, in fact, included a substantial revision in the rate at which the market discounts future profits and thus is a poor guide to the trend of earnings per share.

I conclude that the Princeton formula has more to offer (a) on the grounds of accuracy, despite its conceptual difficulties, and (b) on the grounds of practicality, through its hedge against error via a stabilization fund.

AMERICAN ECONOMIC ASSOCIATION

PROCEEDINGS
OF THE
EIGHTY-SIXTH ANNUAL MEETING

NEW YORK, NEW YORK
DECEMBER 28-30, 1973

THE JOHN BATES CLARK AWARD

Citation on the Occasion of the Presentation of the Medal to

FRANKLIN M. FISHER

December 29, 1973

THE John Bates Clark medal of the American Economic Association is awarded biennially "to that American economist under the age of forty who is adjudged to have made a significant contribution to economic thought and knowledge." For 1973 the Association is proud to award the medal to Franklin M. Fisher, who has contributed continuously and significantly to many branches of theoretical and applied economics, from welfare economics and index number theory to the costing of educational loan systems, and from aggregation theory to the estimation of technological and behavioral relationships in the petroleum drilling and community television antennae industries. In his theoretical work he has concentrated on both supporting the empirical worker in his attempt to wrest meaningful conclusions from recalcitrant data and on keeping the same worker honest by outlining the limits of what is knowable from a given body of data. An example of the first is his work on approximate specification and near decomposability, providing a rationale for much of current econometric practice of stopping short of the estimation of complete systems with potentially an infinity of variables. An example of the second is his contributions to theory of identification of econometric systems and the existence or, more often, nonexistence of aggregates and relationships (such as aggregate production functions) among them. In his empirical work he has lived up to the high standards he set for himself in his methodological contributions. His work represents one of the finest examples of the interaction of theory and measurement in economics.

MINUTES OF THE ANNUAL MEETING
NEW YORK, NEW YORK
DECEMBER 29, 1973

The Eighty-sixth Annual Meeting of the American Economic Association was called to order by President Kenneth J. Arrow in the Grand Ballroom of the New York Hilton Hotel, New York, N. Y., at 9:40 p.m. on December 29, 1973. The President asked that members occupy the center sections of the Ballroom and that nonmembers move to the side sections. He also requested that members offering motions identify themselves by name and give the exact wording of motions to the Secretary in writing.

The Secretary read the minutes of the meeting of December 29, 1972. He pointed out that in accordance with a decision taken at that meeting copies of the reports of the Secretary-Treasurer and the Managing Editors had been made available at a desk in the registration area for the convention before the Annual Meeting and were also distributed at the Meeting itself. In accordance with another decision, microphones were available in the center aisle for the convenience of members wishing to speak. It was VOTED to approve the minutes as read and published in the *American Economic Review,* May 1973, pages 471–73.

The Secretary presented the report of the Committee on Elections and the certification of the new officers for 1973 as follows:

> In accordance with the bylaws on election procedure, I hereby certify the results of the recent balloting and present the results of the Nominating Committee, the Electoral College, and the Committee on Elections.
>
> The Nominating Committee, consisting of Wassily Leontief, Chairman, Carolyn S. Bell, Leonid Hurwicz, Mark Killingsworth, Roy Radner, and David C. Smith, submitted the nominations listed below for Vice-President and member of the Executive Committee. The Electoral College, consisting of the Nominating Committee and the Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.
>
> *President-elect*
>
> Robert Aaron Gordon

Vice-Presidents	*Executive Committee*
Gary S. Becker	Andrew F. Brimmer
Barbara R. Bergmann	Anne Pitts Carter
John G. Gurley	Peter A. Diamond
Geoffrey H. Moore	Paul M. Sweezy

> The Secretary prepared biographical sketches of the candidates and distributed ballots in September. The Committee on Elections, consisting of Stanley W. Black, Chairman, Dawn Elvis, and Rendigs Fels, *ex officio*, canvassed the ballots and filed the results with the Secretary. From the Report of the Committee on Elections, I have the following information:

Number of envelopes without name for identification....	351
Number of envelopes received too late...........	41
Number of defective ballots.	2
Number of legal ballots.....	6,002
	6,396

> On the basis of the canvass of the votes, I certify that the following persons have been duly elected to the respective offices:
>
> *President-elect* (for a term of one year)
>
> Robert Aaron Gordon
>
> *Vice-Presidents* (for a term of one year)
>
> Gary S. Becker
> John G. Gurley
>
> *Members of the Executive Committee*
> (for a term of three years)
>
> Andrew F. Brimmer
> Paul M. Sweezy

The Secretary-Treasurer, the Managing Editor of the *American Economic Review*, and the Managing Editor of the *Journal of Economic Literature* commented on their written reports. (Revised versions of their reports will be published in the *Proceedings* of the Association.) On recommendation of the Treasurer, it was VOTED to adopt the following resolution:

> BE IT RESOLVED that 80 percent of the dues of annual members be considered payment for a year's subscription to both the *American Economic Review* and the *Journal of Economic Literature* and the same dollar amount of the payment by subscribers (nonmembers) be considered payment for subscriptions to the *Review* and the *Journal*, with the remainder a contribution toward the Association's exempt functions, such as annual meetings and various economic research projects and programs.

The Secretary presented the following two resolutions, which were adopted unanimously:

> BE IT RESOLVED that this meeting commend Walter W. Heller for planning a program of great interest and high distinction.

> BE IT RESOLVED that this meeting record a special vote of thanks to the 1973 Local Arrangements Committee headed by Co-chairmen Edward Marcus and John I. Griffin for their hard work and efficient management of the annual meetings of the Allied Social Science Associations.

President Arrow introduced the new President, Walter W. Heller, who took the chair and expressed appreciation for the honor. He announced that his platform had but one plank, to end the meeting by midnight.

Carolyn S. Bell made a report on behalf of the Committee on the Status of Women in the Economics Profession. She called attention to the report of the Committee published in the *American Economic Review*. In reply to a question, she said that she would make recommendations to the Executive Committee in March on whether to extend the life of her Committee beyond May 1975 when its present mandate ends.

F. Ray Marshall made a report on behalf of the Committee on Hiring Practices. He summarized the recommendations the Committee made to the Executive Committee and the action taken on them on December 27, details of which are in the minutes of the Executive Committee to be published in the *Proceedings*.

The President commended the two Committees for their excellent work. Joseph A. Hasson inquired whether the "cartel" would participate in the open listing of job vacancies proposed by the Committee on Hiring Practices. Marshall replied, "We hope so," adding that that intention underlay the resolution proposed by the Committee making it a professional obligation to list vacancies. C. S. Bell suggested that the Association consider a resolution that recommendations not be given for people under consideration for jobs that had not been listed. Charles C. Holt applauded the objectives of the two reports but warned that open listing of jobs might not improve the operation of the job market. It would help neither Harvard nor job applicants, he said, if the listing of a vacancy there drew hundreds of letters from people not qualified for it. Holt urged the Committees to recognize that they were operating in areas in which our knowledge is thin and to experiment with various alternatives. Marshall replied that the advantages of open listing outweigh the disadvantages, especially if the desired qualifications are stated specifically and applicants are advised to apply only for vacancies for which they are qualified. Ali M. S. Fatemi urged that the Association take responsibility for maintaining academic standards and also preserving faculty positions as well as for improving the market for vacancies. The President suggested that he make specific suggestions prior to the meeting of the Executive Committee scheduled for March 8, 1974.

The President expressed appreciation to Rendigs Fels and his staff, particularly Mary Bobo, Administrative Director, and Barbara MacPhee, the former Administrative Director and current Convention Manager.

In calling for new business the President stated that he had received four resolutions before the meeting began, which would be taken up in the order in which he had received them.

Nicholas Georgescu-Roegen read a statement entitled "Toward a Human Economics" and moved that it be published in the *Proceedings*. The motion was seconded by Richard M. Bell. Kenneth Boulding spoke in support of the motion, pointing out that endorsement was not being sought. Robert Eisner said that it was unwise to tell the editor of the *Proceedings* what to publish. R. M. Bell asked if Georgescu-Roegen would accept a change in his motion to make it read "recommend that it [the statement] be published." Georgescu-Roegen accepted the change. Arnold Harberger suggested that the sponsors of the resolution buy space in the *American Economic Review*. William S. Vickrey pointed out that the *Proceedings* differ from other issues of the *American Economic Review*, which are controlled by its Editor; the *Proceedings*, in contrast, are the proceedings of the Association.[1] Alfred Kraessel moved to amend the motion to add the words "and comments and opinions of the members be solicited." The motion to amend was seconded. After some discussion which wandered in imparliamentary fashion from the amendment to the main motion, the previous question was moved and seconded. In response to a parliamentary inquiry, the Secretary explained that moving the previous question was a parliamentary device for ending debate. If passed by a two-thirds majority, it would require proceeding immediately to a vote on the immediately pending business, in this case, the motion to amend. The previous question was put to a vote and passed. The amendment was then put to a vote and passed. After further discussion the previous question was moved, seconded, and passed. At this point the motion as changed and amended was: to recommend that the statement "Toward a Human Economics" be published in the *Proceedings* and comments and opinions of the members be solicited. It was put to a vote and PASSED. The statement appears below as an appendix to these minutes.[2]

---

[1] James Duesenberry in a private conversation during the meeting pointed out to the Secretary that, since Georgescu-Roegen had read the statement, it would *ipso facto* appear in the *Proceedings*. He felt that further discussion of the motion and the proposed amendment was unnecessary. The Secretary appreciated the irony that would arise if the motion lost and he nevertheless included the statement in the minutes. Relishing irony, he was tempted to speak against the motion. But sober reflection on the policy governing minutes previously adopted by the Executive Committee on his recommendation* convinced him that he ordinarily has discretion as to whether to include in the minutes a long statement read at the meeting. The test of what to include in the minutes is what the members will later want to read. Whether this footnote could survive the test is questionable, but clearly (1) defeat of the main motion would have indicated the members did not want to read the statement in the *Proceedings* and (2) R. M. Bell's change preserved the flexibility of present policy toward the minutes. Cf. footnotes 2, 3, 4 and 6 below.

> * "The principal object of the minutes is to record actions taken. Inclusion of anything more . . . should be governed by the sole criterion of whether enough people will want to read it to make inclusion worth the cost . . . editorial judgment must be exercised, normally by the Secretary, but occasionally by the Executive Committee when its judgment differs from his. Judgment calls are likely to be disputed by those adversely affected, as every umpire and referee knows, but that is no reason for shrinking from making them." (Minutes of the Executive Committee.)

[2] The policy quoted in the footnote to the previous footnote clearly calls for publication even though the motion itself gave the Secretary discretion as to whether to do so.

Mr. David Laibman moved adoption of the following resolution:[3]

> WHEREAS the constitutionally elected government of Chile was overthrown by a military coup d'etat on September 11, 1973;
>
> WHEREAS the junta currently in power appears to have indefinitely suspended the Chilean Congress, destroyed the independence and intellectual autonomy of the universities, and persecuted and imprisoned former members and supporters of the Allende Government;
>
> AND WHEREAS there are reported to be Chilean professionals, economists included, in concentration camps, in hiding, and in foreign embassies awaiting exit visas (in effect imprisoned by the junta);
>
> THEREFORE BE IT RESOLVED THAT the American Economic Association goes on record as expressing concern over reports of suppression of academic freedom and the persecution of economists and other professionals in Chile;
>
> AND BE IT FURTHER RESOLVED THAT the American Economic Association calls upon economists in the United States to express support for their Chilean colleagues; to seek means of offering employment to Chilean economists in *U.S.* colleges and research institutes; and to register their concern to *U.S.* and Chilean governmental agencies in ways they deem appropriate.

The motion was seconded. The previous question was moved and seconded but failed to gain the necessary two-thirds majority. After some discussion, George Stigler stated that the resolution violated the charter of the Association. The President called on the Secretary, who read the section of the Certificate of Incorporation on the objectives of the Association: "The particular business and objects of the society are as follows: 1. The encouragement of economic research, especially the historical study of the actual conditions of industrial life; 2. The issue of publications on economic subjects; 3. The encouragement of perfect freedom of economic discussion. The Association as such will take no partisan attitude, nor will it commit its members to any position on practical economic questions." The President then ruled the resolution in order. After further discussion, the previous question was again moved and seconded. The hour being late, it now passed. The main motion was then defeated.

It was moved, seconded, and passed that speeches be limited to three on each side and five minutes per speech.[4]

Richard England[5] read a resolution on racism and moved its adoption. The President ruled it out of order on grounds that it was inconsistent with the provision in the Certificate of Incorporation forbidding the Association to take a position on political questions. The ruling was appealed, and the appeal seconded. The President ruled that the appeal was debatable. Because this latter ruling was disputed, the Secretary read the relevant portion of *Robert's Rules of Order* (see Henry M. Robert). After some debate, the ruling of the President that the resolution on racism was out of order was sustained.[6]

It was moved and seconded that the

---

[3] The Secretary has exercised his discretion to include the text of the resolution in full, partly because it is not excessively long and partly because the ruling by the chair that it was in order (see below) established a precedent that may be of future interest.

[4] The Secretary interprets this motion as applying only to the meeting of December 29, 1973, and not to subsequent meetings.

[5] After the meeting the files of the Association were searched in vain for a member in good standing of this name. A Richard W. England had been dropped from membership for nonpayment of dues.

[6] The Secretary has exercised the discretion described in footnote 1 to omit the text of the resolution on racism because (1) it is excessively long, (2) it is similar to a shorter resolution offered at the meeting of December 1972, which was ruled out of order* but nevertheless was published in the *Proceedings*, and (3) to publish any statement of any length that is offered as a resolution, even though out of order, would provide undue temptation to members of a profession under mandate to publish or perish.

* On appeal, duly seconded, the ruling was upheld by vote of the members present. See Minutes of the Annual Meeting.

meeting be adjourned. The President pointed out that he had been given another resolution before the meeting that had not been acted on, but conceded that a motion to adjourn was not debatable. A member inquired what the resolution was. The President asked the Secretary whether it would be fair to say. The Secretary replied, "Fair but unparliamentary." The motion to adjourn was withdrawn. The President then revealed that the remaining resolution provided for the Annual Meeting henceforth to be scheduled before the Presidential Address rather than afterwards to avoid meeting far into the night. At a time declared by the President to be 11:59 p.m., it was moved, seconded, and passed that the meeting be adjourned, thus (1) enabling the President to make good on the only plank in his platform, (2) showing that the members are willing to meet till midnight but not beyond, and (3) suggesting that the President's watch needed adjustment.

RENDIGS FELS, *Secretary*

APPENDIX

Toward a Human Economics[7]

The evolution of our global household earth is approaching a crisis on whose resolution man's very survival may depend, a crisis whose dimensions are indicated by current rates of population expansion, runaway industrial growth, and environmental pollution, with their attendant threats of famine, war, and biological collapse.

This evolution, however, has not been determined solely by inexorable laws of nature, but by the human will operating within nature. Man has shaped his destiny through a history of decisions for which he is responsible; he can change the course of that destiny by new conscious decisions, by a new exertion of will. To begin with, he requires a new vision.

Basic to our function as economists is the description and analysis of economic processes as we observe them in operation. Increasingly over the last two hundred years, the economists have been called upon, and have undertaken, not merely to analyze, theorize, describe, and measure the economic scene, but also to advise, to plan, and to take an active part in the conduct of affairs. The power of the economists, and therewith their responsibility, has become very great indeed.

In the past, production has been regarded as a benefit. But production also entails costs that have only recently become apparent. Production necessarily drains our finite stock of raw materials and energy, while it floods the equally finite capacity of our ecosystem with the wastes of its processes. The economist's traditional measure of national and social health has been growth. But continued industrial growth in areas already highly industrialized is a short-term value only; present production continues to grow at the expense of future production, and at the expense of the delicate and evermore threatened environment.

The reality that our system is finite and that no expenditure of energy is free, confronts us with a moral decision at every point in the economic process, in planning and development and production. What do we need to make? What are the real, long-term costs of production, and who is required to pay them? What is truly in the interests of man, not in the present only, but as a continuing species? Even the clear formulation from the economist's perspective of the choices before us is an ethical task, not a purely analytical one, and economists ought to accept these ethical implications of their work.

We call upon our fellow economists to embrace their role in the management of our earth home, and to join the efforts of other scientists and planners indeed, of other men and women in all areas of thought and endeavor, to ensure the survival of man. The science of economics, like other fields of inquiry in search of precision and objectivity, has tended in the last century increasingly to isolate its domain from others. But the time when economists could fruitfully work in isolation is gone.

---

[7] Comments and opinions on this statement may be sent to the Secretary, American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212.

We must have a new economics whose purpose is the husbanding of resources and the achievement of rational control over the development and application of technology to serve real human needs rather than expanding profits, warfare, or national prestige. We must have an economics of survival, still more, of hope—theory and vision of a global economy based on justice, which would make possible the equitable distribution of the earth's wealth among its people, present and future.

It is clear that we can no longer usefully consider apparently separate national economies apart from their relations to the larger global system. But economists can do more than measure and describe the complex interrelations among economic entities; we can work actively for a new order of priorities that transcends the narrow interests of national sovereignty and serves instead the interests of the world community. We must replace the ideal of growth, which has served as a substitute for equitable distribution of wealth, with a more humane vision in which production and consumption are subordinated to the goals of survival and justice.

Currently, a minority of the earth's people enjoy an inordinate share of resources and industrial capacity. These industrial economies, capitalist and socialist alike, must find ways to cooperate with developing economies to correct the imbalance, without pursuing ideological or imperialist competition, and without exploiting the people they propose to aid. In order to achieve equitable distribution of wealth throughout the world, the people of the industrialized countries must relinquish what now seems an unbounded right to consume whatever resources are available to them, and we as economists must play a role in the reshaping of human values toward this end. The accidents of history and geography must no longer serve as rationale for injustice.

The task for economists is therefore an extremely novel and difficult one. Many people now look at the available data—the trends of population growth, pollution, resource depletion, and social upheaval—and lose hope. We have already passed the point of no return toward our rendezvous with disaster, they say gloomily; nothing can be done. But despair is a position we must reject. The moral imperative is for us to create a new vision, to make a road to survival through a treacherous country where there are no roads. At the present moment, man possesses the wealth and the technology not only to save himself for a very long future, but to make for himself and for all his children a world in which it is possible to live with dignity and hope and comfort; but he must decide to do it. We call on economists to join in framing the new vision that will enable man to use his wealth in his own interests, disagreeing, perhaps, on details of method and policy, but agreeing emphatically on the goals of survival and justice.

DAI DONG, a project of the
Fellowship of Reconciliation
Box 271, Nyack, N.Y. 10960
Hasselhaven 6, 2600 Glostrup, Denmark
October, 1973

REFERENCES

"Charter and Bylaws of the American Economic Association," *1969 Handbook of the American Economic Association, Amer. Econ. Rev.*, January 1970, *59*, v.

Committee on the Status of Women in the Economics Profession, "Combatting Role Prejudice and Sex Discrimination," *Amer. Econ. Rev.*, Dec. 1973, *63*, 1049–61.

"Minutes of the Annual Meeting of the American Economic Association, December 27, 1972," *Amer. Econ. Rev., Proc.*, May 1973, *63*, 473.

"Minutes of the Executive Committee," *Amer. Econ. Rev., Proc.*, May 1971, *61*, 478.

H. M. Robert, *Robert's Rules of Order Newly Revised*, a New and Enlarged Edition by Sarah Corbin Robert, Glenview, Ill., 1970, 220.

MINUTES OF THE EXECUTIVE COMMITTEE MEETINGS

Minutes of the Meeting of the Executive Committee in New York City, March 2, 1973.

The *first meeting* of the 1973 Executive Committee of the American Economic Association was called to order in the New York Hilton Hotel, New York, N. Y., at 9:45 a.m. on March 2, 1973. The following members were present: Kenneth J. Arrow, presiding, Irma G. Adelman, George H. Borts, James S. Duesenberry, Robert Eisner, Rendigs Fels, Robert Heilbroner, Walter W. Heller, Lawrence R. Klein, John R. Meyer, Guy H. Orcutt, Joseph A. Pechman, and Mark Perlman. Absent were John Kenneth Galbraith and James Tobin. Present as members of the Nominating Committee were: Wassily Leontief, Chairman, Carolyn Shaw Bell, Leonid Hurwicz, Mark R. Killingsworth, Roy Radner, and David C. Smith. Robert Solow was present as Chairman of the Committee on Honors and Awards, and Ray Marshall was present as Chairman of the Committee on Hiring Practices.

1. *Minutes.* The minutes of the meeting of December 27, 1972, were approved.

2. *Report of the Secretary* (Rendigs Fels). The Secretary reported the results of his investigations of possible meeting sites for 1979 and 1980. It was VOTED to meet in Atlanta, December 28–30, 1979, and to meet in Denver during the week of September 2–6, 1980. The Secretary was asked to investigate Washington and New Orleans as possible meeting sites for 1981. It was VOTED to ask him to investigate the possibility of changing the site of the 1976 meetings to Philadelphia. He reported that several nonmembers participated in the Annual Meeting held December 29, 1972. It was VOTED that at future annual meetings the President ask nonmembers to sit in side sections or balconies with the center section of the meeting room reserved for members, that each person who speaks be required to identify himself so that the Secretary can record his name, and that applications for membership be accepted at the meeting itself. It was VOTED that convention hotel rates be made available only to those who preregister for the meetings. The Secretary announced his intention to discontinue publishing abstracts of the papers presented at the meetings; there was no objection. It was VOTED to authorize the Secretary to hire a Convention Manager. It was VOTED to establish the position of Assistant Secretary-Treasurer to be filled by appointment of the President on nomination by the Secretary-Treasurer.

3. *Report of the Treasurer* (Rendigs Fels). The Treasurer presented preliminary financial reports prepared by the auditors. They were received too late to make possible preparation of a revised budget for 1973. The Treasurer presented evidence that relevant cost and price indexes will have risen considerably more than 5 percent from the end of 1971 to the end of 1973. Under the authority of the bylaws (Section I, Paragraph 2) it was VOTED to increase dues 5 percent effective January 1, 1974, subject to reconfirmation at the meeting of the Executive Committee on December 27, 1973. It was VOTED to authorize the Treasurer to change the fiscal year. It was agreed to authorize the Secretary-Treasurer to appoint advisers on problems of computerizing and integrating membership records, the handbook, and bibliographical information generated by the *Journal of Economic Literature.*

4. *Report of the Editor of the American Economic Review* (George H. Borts). The Managing Editor reported that he had instituted the policy of anonymous refereeing of every other manuscript, as decided at the Annual Meeting of December 29, 1972.

5. *Report of the Editor of the Journal of Economic Literature* (Mark Perlman). The Managing Editor reported that the number of journals indexed is continuing to increase. In view of the difficulties experienced, he is considering abandonment of computerization.

6. *Report of the Editor of the Papers and Proceedings* (Rendigs Fels). The Editor submitted a written report detailing the problems he has encountered. It was agreed that henceforth papers exceeding the space limit specified by the Program Chairman would not be published in the *Papers and Proceedings* but would instead be referred to the Managing Editor of the *American Economic Review* and put through the normal refereeing procedures. The President reiterated his intention to appoint a new committee on the annual meetings.

7. *Committee on Honors and Awards* (Robert M. Solow). The Chairman submitted the recommendations of the Committee on Honors and Awards. The Electoral College consisting of that Committee and the Executive Committee meeting together VOTED unanimously to award the John Bates Clark Medal to Franklin Fisher.

8. *Nominating Committee* (Wassily Leontief). The Chairman reported the following nominees for offices in the 1973 election: Vice-President (two to be chosen), Gary Becker, Barbara Bergmann, John G. Gurley, and Geoffrey H. Moore; members of the Executive Committee (two to be chosen), Andrew Brimmer, Anne P. Carter, Peter A. Diamond, and Paul M. Sweezy. The Electoral College consisting of the Nominating Committee and the Executive Committee meeting together chose as nominee for President-elect Robert Aaron Gordon, as Distinguished Fellow Tibor Scitovsky, and as representative to the Social Science Research Council Lawrence R. Klein.

9. *Committee on Hiring Practices* (F. Ray Marshall). The Chairman proposed publication of a pamphlet on the labor market for economists, to include projections of supply and demand for economists, information about how the market works, advice to applicants, recommendations to economics departments about placement of graduate students, and recommendations to prospective employers on such matters as how to conduct interviews. It was agreed that the Committee be expanded to include two additional members and that it would make a report to the 1973 Annual Meeting. It was VOTED to approve the proposed manual and to approve search by the Committee for foundation support for a research study.

10. *Committee on the Status of Women in the Economics Profession* (Carolyn Shaw Bell). The Chair requested $7,500 to complete work already well advanced on a roster of women economists. After discussion, it was VOTED to appropriate $5,000 for the Secretary's office to carry out this task in conjunction with the Committee.

11. *Program Committee* (Walter W. Heller). The Chairman announced that the 1973 program will emphasize public policy, particularly macroeconomics. All sessions will be based on invited papers. Efforts are being made to bring foreign economists and officials to participate in the program.

12. *Committee on Policy Communication* (Rendigs Fels for Frank Schiff). The Secretary reported that the sample Washington Newsletter published in the December 1972 issue of the *American Economic Review*, pages 1024–28, had resulted in expressions of interest by less than twenty

economists, indicating that there was too little support for the Newsletter to warrant proceeding with regular publication. It was agreed to discharge the Committee with thanks for its efforts.

13. *Committee on Economic Education* (Rendigs Fels for G. L. Bach). A written report was submitted giving details of the series of readings books proposed at the previous meeting of the Executive Committee. The series was approved with the understandings that there would be two editors, that the senior editor would receive an honorarium of $1,000 per volume, that the junior editor would receive one-ninth of his annual salary, and that the publisher would advance the fees for reprint rights.

14. *University of Rochester* (Richard N. Rosett). A letter to the President from Professor Richard N. Rosett of the University of Rochester was distributed and discussed. The letter gave the reasons for refusal to respond to the questionnaire of the Committee on the Status of Women in the Economics Profession.

15. *Counsel.* It was VOTED to pay the Counsel a retainer of $1,000 a year.

16. *Expiration of Terms of Editors.* In view of the fact that the terms of office of the managing editors of the *American Economic Review* and the *Journal of Economic Literature* expire in 1974, the President announced that he would appoint a special committee to make recommendations at the next meeting of the Executive Committee.

17. *Date of Spring Meeting.* It was agreed that the Executive Committee would meet on March 8–9, 1974, in Washington, D.C.

The meeting adjourned at 6:05 p.m.

Minutes of the Meeting of the Executive Committee in New York City, December 27, 1973.

The second meeting of the 1973 Executive Committee was called to order at 2:15 p.m. on December 27, 1973, in the New York Hilton Hotel, New York, N.Y. The following members were present: Kenneth J. Arrow, presiding, Irma G. Adelman, George H. Borts, James S. Duesenberry, Robert Eisner, Rendigs Fels, John Kenneth Galbraith, Robert L. Heilbroner, Walter W. Heller, Lawrence R. Klein, John R. Meyer, Guy Henderson Orcutt, Joseph A. Pechman, Mark Perlman, and James Tobin. Also present were newly elected members Andrew F. Brimmer, Robert Aaron Gordon, and John G. Gurley. Absent were newly elected members Gary S. Becker and Paul M. Sweezy. Present as Counsel was Donald F. Turner. Present as guests for parts of the meeting were Carolyn S. Bell, William V. Brainard, Robert Dorfman, Robert Ferber, Gary Fromm, Fritz Machlup, F. Ray Marshall, and Charles A. Taff.

1. *Minutes.* The minutes of the meeting of March 2, 1973, were approved as circulated.

2. *Report of the Secretary* (Rendigs Fels). A written report by the Secretary had been sent to members of the Executive Committee in advance of the meeting giving information about future meeting sites and other matters. It was VOTED that the 1981 meetings be in Washington, D.C., during the Christmas holidays.

3. *Report of the Treasurer* (Rendigs Fels). The Treasurer commented on the report previously distributed giving revised figures for the expected results for 1973 and the proposed budget for 1974. In separate actions, it was VOTED to publish a new *Handbook* in 1974 or as soon thereafter as possible, to reconfirm the dues increase of 5 percent approved at the meeting of March 2, 1973, and to approve the revised budget for 1974 with an estimated deficit of $14,000.[1]

---

[1] Details of the revised budget will be given in the Treasurer's Report in the *Proceedings*.

4. *Report of the Managing Editor, American Economic Review* (George H. Borts). The Editor amplified on his written report. He announced his intention of going over completely to anonymous refereeing. He will discontinue publishing abstracts of Ph.D. dissertations in view of the fact that this service duplicates one already available from another source. There was considerable discussion of the editorial policies of the *Review*. It was VOTED to approve as new members of the Board of Editors Martin Feldstein, Bent Hansen, Jerome Stein, S. C. Tsiang, Finis Welch, and Marina v. N. Whitman. It was VOTED to ask the Committee on the Journals to investigate the possibility of some relation between the Association and *Challenge* magazine and to report to the next meeting of the Executive Committee.

5. *Report of the Managing Editor, Journal of Economic Literature* (Mark Perlman). The Editor commented on his written report. It was VOTED to approve the appointment as members of the Board of Editors Tibor Scitovsky, Anne P. Carter, and William S. Vickrey.

6. *Committee on the Journals* (William C. Brainard). On recommendation of the Committee it was VOTED to reappoint Borts and Perlman to additional three-year terms as Managing Editors of the *Review* and the *Journal* respectively.

7. *Report of the Counsel* (Donald F. Turner). The Counsel discussed the provision of the bylaws requiring that all publications of the Association be distributed to members. He interpreted it to include only the *Review*, the *Journal*, and the *Handbook*. He felt that a clarifying change to exclude books was not needed. It was VOTED to accept the recommendation of the Counsel. (This action was superseded later in the meeting. See paragraph 12 below.)

8. *Fulbright-Hayes Fellowships* (Irma G. Adelman). Adelman discussed the problem of getting better applicants for Fulbright-Hayes fellowships. The President appointed her a committee of one to draw up an appropriate resolution for consideration at the March meeting of the Executive Committee.

9. *Intersociety Committee on Transportation* (Charles A. Taff). Taff reported that a new constitution had been drawn up for the Intersociety Committee on Transportation. The text was not available but, insofar as known, its provisions were acceptable, including those governing the financial obligations of the constituent societies. It was VOTED to approve the constitution subject to review by the Secretary.

10. *Journal of Research on Consumer Behavior* (Robert Ferber for Mary Jean Bowman). On Ferber's recommendation, it was VOTED to permit free use of the mailing list and free advertising for soliciting subscriptions to the new *Journal of Research on Consumer Behavior.*

11. *International Economic Association* (Fritz Machlup). Several matters relating to the International Economic Association were discussed. It was agreed to authorize the incoming President to proceed with discussions designed to promote colloquia of Russian and American economists. For this purpose he may appoint a committee and will work closely with Machlup and Meyer.

12. *Committee on Hiring Practices* (F. Ray Marshall). A discussion paper by Kathleen Brook and Marshall entitled "The Labor Market for Economists" had been distributed to the Executive Committee before the meeting. The last section of the paper contained the recommendations of the Committee on Hiring Practices. In view of the Committee's recommendation (1) on pages 56–57 of the Report for open listing of vacancies in a

bimonthly publication with a subscription price of, say, $20, it was VOTED to ask the Counsel to draft an amendment to the by-laws to specify that members are entitled to receive the *Review*, the *Journal*, and the *Handbook*, leaving to the Executive Committee discretion to charge for other publications. It was understood that this amendment would be submitted to the membership for ratification by mail ballot. It was then VOTED to approve the substance of recommendation #1, including specifically the proviso that "all members of the *AEA* have a professional obligation to list their openings with the Association," and to ask the Committee to bring to the next meeting of the Executive Committee a detailed plan for implementation of the recommendation. It was then VOTED to approve recommendations (2), (3), and (4) on pages 57–59 of the report. It was VOTED to approve the further recommendations on page 59 of the report, to ask the Secretary to mail to chairmen of economics departments the recommendations on page 59 of the report with notice that they have been approved by the Executive Committee, and to publish same in the notes section of the *Review* together with the sentence quoted above from recommendation (1) and the recommendation that placement officers hold a meeting in conjunction with the annual meetings of the Association. (Pages 56–59 of the report are reproduced at the end of these minutes.)

13. *Committee on the Status of Women in the Economics Profession* (Carolyn Shaw Bell). Bell gave an oral report on the activities of the Committee. The Committee was out of money and was applying to the Ford Foundation for another grant. It was VOTED to appropriate $5,000 to carry on the Committee's work.

14. *Advisory Committee on Minority Group Economists* (Robert Dorfman). The chairman reported that little progress had been made and suggested that the Executive Committee might need to reconsider its strategy. It was agreed to defer any decision until the March meeting.

15. *Census Advisory Committee* (Gary Fromm). Fromm noted that, despite past misgivings of the Executive Committee about the appointment of the new Director of the Bureau of the Census, Vincent P. Barabba, he and the Census Advisory Committee found Barabba dedicated to improving the already high professional standards of the Bureau and to shielding it from political influences. He then read a report on the confidentiality of census statistics and suggested that the Association pass a resolution on the subject. After considerable discussion, he was asked to draft such a resolution and bring it to the March meeting. He also suggested, because the Committee's charter is to be renewed in January 1975, that the Association appoint a committee "to reconsider its role with respect to sponsorship and appointment of members of the Census Advisory Committee, the role and effectiveness of the Committee in advising the U.S. Census Bureau, and its responsibilities in informing the Association's membership and mobilizing support for vital statistical programs." It was agreed to refer this suggestion to the incoming President.

16. *Visiting Economics Scholars Program.* A written proposal for a Visiting Economics Scholars Program by Edward I. Powers had been distributed in advance of the meeting. It would be modeled on the Visiting Scientists Program formerly funded by the National Science Foundation (see the final report by Phillip Saunders). It was VOTED to authorize the Secretary to spend $7,500 on such a program. The discussion indicated that the visiting scholars need not be paid an honorarium. It also indicated need for pre-

cautions to insure that a college getting the benefit of a visiting scholar at nominal cost to itself was not simultaneously paying substantial honoraria to other visiting speakers.

17. *Harold Hotelling.* It was agreed that the President would send a message of condolence to the widow of Harold Hotelling which would express the appreciation of the Association for his contributions to economics.

The meeting adjourned at midnight.

RENDIGS FELS, *Secretary*

Recommendations of the Committee on Hiring Practices

(Excerpted from Brook and Marshall, pp. 56–59. Page numbers in square brackets refer to the pagination of the original report.)

[p. 56] We . . . recommend that the American Economic Association:

(1) a. Collect and publish six times a year (February, April, June, August, October, and December) lists of job openings as well as lists of academic institutions with no openings. The cost of this service could be defrayed by subscription fees of, say, $20 per year. We should resolve that all members of the *AEA* have a professional obligation to list their openings with the Association.

[p. 57] b. Conduct an annual labor market survey each September, to be published in December, perhaps with the December job list. This survey would include an estimate of the numbers of people expecting to enter the labor market and therefore provide some indication of the demand and supply relationship.

c. Explore with the U.S. Employment Service the possibility of establishing a cooperative arrangement to facilitate the distribution of labor market information for and about economists.

d. Establish cooperative arrangements with such organizations as the National Association of Business Economists, the American Bankers Association, and government employers of economists, in order to acquire more information about the nonacademic demand for economists. These nonacademic sources become particularly important as academic job opportunities decline relative to the supply.

The publication of lists of jobs and applicants would make it possible to discontinue listing labor market information in the *American Economic Review,* thus freeing valuable space for other uses. There also are indications that very little use is made of the *AER* listings.

(2) Precede the annual meetings by one day for hiring transactions. There have been numerous complaints about confusion when department chairmen and others interested in interviewing candidates are attempting to do so while programs are in progress. The labor market could be more [p. 58] orderly and those who wished to do so could attend more sessions if the first day were set aside for labor market transactions. Since programs would not be in session, better facilities for interviewing could be made available during the time reserved for hiring. Of course, hiring transactions would continue during the other days of the annual meetings, but hopefully at a less frantic pace than formerly if one day were devoted exclusively to this purpose.

It has sometimes been suggested that separate meetings be held for hiring purposes at some time during the year so as not to conflict with the annual meetings. However, this suggestion seems impractical, and setting aside a day during the annual meetings seems more appropriate.

(3) Make sure that adequate interviewing and communications facilities are available during the annual meetings. It would be particularly useful to provide effective communications systems for department chairmen or their representa-

tives. The local arrangements committee should pay particular attention to planning the operation of labor market information facilities. If suitable arrangements cannot be made in the convention hotel, separate facilities should be established.

(4) Maintain a current list of department chairmen to be sent to participating departments and others on request. This would facilitate the communications process. Departments could send this information to the *AEA* along [p. 59] with news currently collected for announcements sections in the *AER*.

The American Economic Association could also encourage departments to take some actions to improve their placement and hiring procedures. These include:

(1) Recognizing a responsibility to help place students, including appointment of placement officers with adequate time and assistance to stay abreast of market developments.

(2) Counseling graduate students early about the present and probable future state of the labor market and how to behave in the labor market. Advisers should provide realistic information to students about their prospects.

(3) Providing full and accurate information to prospective employers. It is important to anticipate the kinds of data that would be useful to particular employers, with documentary evidence of teaching and research ability wherever available. It also is important not to describe the personal characteristics of women.

With respect to hiring, employers should:

(1) Wherever possible, make appointments with applicants in advance of the annual meetings.

(2) Exercise common courtesies in dealing with applicants. It is inexcusable for employers not to be open with applicants concerning their prospects, to interview several candidates at one time, not to give timely notification of the applicant's status, or to ask questions of women not asked of men.

REFERENCES

K. Brooks and F. R. Marshall, "The Labor Market for Economists," mimeo., 1973 (printed below, pp. 488–511).

P. Saunders, "Report on the Visiting Scientists Program in Economics," *Amer. Econ. Rev., Proc.,* May 1973, *63,* 512.

REPORT OF THE SECRETARY FOR 1973

Annual Meetings. The final report for the annual meetings of the Allied Social Science Associations in Toronto, Ontario, December 28–30, 1972, indicated a total registration of 6,119, including 3,474 members of the American Economic Association. The net income from the meetings was $11,488.87. The share of the Association was $7,490.74. During 1972 the Association incurred out-of-pocket expenses for the Toronto meetings and the planning of later meetings of $6,661. These figures do not include any allowance for the time of Association employees or overhead expenses. Everything considered, the Association derived a positive but small financial return from the meetings.

The final report for the annual meetings held in New York, N. Y., December 28–30, 1973, is not yet available. A total of 6,832 persons used 4,594 hotel rooms obtained through the Allied Social Science Associations.

Until now, the local arrangements committees have had a heavy burden, borne almost entirely by volunteer workers. In the future, the management of the meetings will be increasingly professionalized. As the first step in this direction, Barbara MacPhee has been appointed Convention Manager.

The schedule for future meetings is: December 28–30, 1974, San Francisco, San Francisco Hilton; October 2–4, 1975, Dallas, Dallas Hilton; September 16–18, 1976, Atlantic City, Chalfonte-Haddon Hall; December 28–30, 1977, New York, New York Hilton; August 29–30, 1978, Chicago, Conrad Hilton; December 28–30, 1979, Atlanta, Atlanta Hilton and Peachtree Center Plaza; September 2–6, 1980, Denver, Denver Hilton; December 28–30, 1981, Washington, D. C., Washington Hilton and Sheraton-Park.

TABLE 1—MEMBERS AND SUBSCRIBERS

(End of Year)

	1970	1972	1973
Class of Membership			
Annual	16,075	14,462	14,453
Junior	2,077	1,886	2,480
Family	289	262	277
Complimentary	160	345	375
Life	291	315	332
Honorary	16	72	72
Total Members	18,908	17,342	17,989
Subscribers	7,265	7,147	7,228
Total Members and Subscribers	26,173	24,489	25,217

Membership. Table 1 shows that the total number of members and subscribers in December 31, 1973, was 25,217, a net increase of 728 over December 31, 1972. This small rise follows a two-year decline from the peak of 26,173 reached at the end of 1970. The decline resulted from doubling the dues and tripling the price of subscriptions. Demand by subscribers appears to be less elastic than demand by members. In spite of the much larger increase in the price of subscriptions, the number of subscribers is less than 1 percent below the 1970 peak, whereas the number of members is 5 percent lower. Moreover, there has been a change in the composition of the membership. The number of junior members (who pay only half the dues of annual members) has increased 19 percent, whereas the number of annual members has declined 10 percent.

Of the 17,989 members at the end of 1973, 14,878 resided in the United States, 1,025 in Canada, and the remaining 2,086 in other foreign countries. Whether the membership includes most or less than half the number of economists in the United States depends on how the term is defined. The journals and other services of the Association appear to have great value for economists engaged in research and university teaching and who have the Ph.D. or equivalent training. This emphasis is in accordance with the objectives of the Association as stated in its Certificate of Incorporation. The survey articles in the *Journal of Economic Literature*, the special articles on the Economic Report of the President in the *American Economic Review*, and the expository articles in the *Papers and Proceedings* go a considerable way toward meeting the needs of other economists. The question remains, though, as to whether the Association should broaden its efforts.

Handbook. The Executive Committee on December 27, 1973, voted $150,000 for a new edition of the *Handbook*. Like the 1969 edition, the new one will include biographical and bibliographical information. Questionnaires will be sent to members this spring. The *Handbook* will be distributed free to members.

Employment Services. Until now, three types of employment services have been available to economists. One is the "Vacancies and Applications" section of the *American Economic Review*. The second is the National Registry for Economists, operated on a year-round basis by the Chicago Placement Service of the Illinois State Employment Service under the direction of Mrs. Theresa Scholl. The third is the employment service provided at the annual meetings of the Allied Social Science Associations by the State offices of the U. S. Employment Service. See the Minutes of the Executive Committee meeting of December 27, 1973, (above) and the Report of the Committee on Hiring Practices (below) for a description of the additional services now in the planning stage.

Insurance. During 1973, a group life insurance plan was made available to members through the Smith-Sternau Organization, Inc. More than a thousand members are participating.

Permission to Reprint and Translate. Official permissions to quote from, reprint, or translate and reprint articles from the *American Economic Review* and the *Journal of Economic Literature* totaled 147 in 1973 compared to 186 in 1972 and 230 in 1971. Upon receipt of a request for permission to reprint, the publisher or editor is instructed to get the author's permission in writing and to send a copy to the Secretary as a condition for official permission. The Association suggests that the author charge a fee of $150, but he may charge some other amount, enter into a royalty arrangement, or waive the fee.

Visiting Economics Scholars Program. As recorded above in the minutes of the meeting held on December 27, 1973, the Executive Committee has decided to finance a Visiting Economics Scholars Program modeled on the Visiting Scientists Program in Economics formerly supported by the National Science Foundation. The new program will be administered by the Secretary. Its purpose is to facilitate visits by leading economists to smaller colleges emphasizing undergraduate teaching. The colleges will be expected to pay part or all of the costs of the visits in accordance with their financial resources; at a minimum, they will take care of the local expenses of the visitor. Inquiries should be directed to the Secretary.

Standing Committees

ADVISORY COMMITTEE ON THE HISTORY OF THE ASSOCIATION

George J. Stigler, *Chairman*
Joseph Dorfman
Harold F. Williamson, *Corresponding Secretary*

COMMITTEE ON ECONOMIC EDUCATION

G. L. Bach, *Chairman* (1976)
Elisabeth Allison (1975)
Kenneth E. Boulding (1975)
Allen C. Kelley (1975)
Phillip Saunders (1974)
Henry H. Villard (1973)
Rendigs Fels, *Ex Officio*

COMMITTEE ON HONORS AND AWARDS

Robert M. Solow, *Chairman* (1974)
Irma Adelman (1978)
Marcus Alexis (1978)
Gary Becker (1976)
Lloyd G. Reynolds (1974)

BUDGET COMMITTEE

Joseph A. Pechman, *Chairman*
Irma Adelman
Rendigs Fels, *Ex Officio*
Kenneth J. Arrow, *Ex Officio* (1973)
Walter W. Heller, *Ex Officio* (1974)

FINANCE COMMITTEE

Beryl W. Sprinkel, *Chairman*
C. Wells Farnham
Rendigs Fels
Milton Friedman
Walter W. Heller

CENSUS ADVISORY COMMITTEE

Gary Fromm, *Chairman* (1973)
Charles Holt (1973)
John Kendrick (1973)
Murray Weidenbaum (1973)
Richard Wirthlin (1973)
Jacob Mincer (1974)
Guy E. Noyes (1974)
Lee E. Preston, **Chairman* (1974)
Joseph Burns (1975)
Anne P. Carter (1975)
Zvi Griliches (1975)
Dean E. McKee (1975)
Richard Ruggles (1975)
J. Fred Weston (1975)
*Gardner Ackley (1976)
*Armen Alchian (1976)
*Andrew F. Brimmer (1976)
*Anthony Downs (1976)
*James R. Nelson (1976)

ECONOMICS INSTITUTE, POLICY AND ADVISORY BOARD

Henry Rosovsky, *Chairman* (1974)
Albert Fishlow (1973)
Earl O. Heady (1973)
James Worley (1974)
Irma G. Adelman (1975)
Edwin S. Mills, **Chairman* (1975)
Walter P. Falcon (1976)
Daniel Schydlowsky (1976)

JOINT COMMITTEE WITH THE ASSOCIATION OF AMERICAN LAW SCHOOLS

Richard A. Musgrave, *Co-Chairman* (1973)
Louis De Alessi (1974)
Harold Demsetz (1974)
*Alvin Klevorick (1976)
Gerald M. Meier (1975)
Peter O. Steiner (1975)
George J. Stigler, **Co-Chairman* (1976)

---

* Newly appointed in 1974.

Special Committees

COMMITTEE ON HIRING PRACTICES

F. Ray Marshall, *Chairman*
Carolyn Shaw Bell
Robert J. Brown
Allan Cartter
Stephen Marglin

NOMINATING COMMITTEE (1973)

Wassily Leontief, *Chairman*
Carolyn S. Bell
Leonid Hurwicz
Mark Killingsworth
Roy Radner
David C. Smith

NOMINATING COMMITTEE (1974)

James Tobin, *Chairman*
Allen C. Kelley
George L. Perry
Roy Radner
Christopher A. Sims
Alice Vandermeulen
Charles Z. Wilson

COMMITTEE ON ELECTIONS (1973)

Stanley W. Black, *Chairman*
Dawn Elvis
Rendigs Fels, *Ex Officio*

ADVISORY COMMITTEE ON MINORITY GROUP ECONOMISTS

Robert Dorfman, *Chairman*
Marcus Alexis
George H. Borts
Andrew Brimmer
Karl Gregory
Charles Schultze
Thaddeus Spratlen
James Tobin

INSURANCE COMMITTEE

Arthur M. Okun, *Chairman*
John Lehman
Joseph A. Pechman

COMMITTEE ON THE JOURNALS

William Brainard, *Chairman*
Mary Jean Bowman
Robert Fogel
John G. Gurley
Burton A. Weisbrod

COMMITTEE ON THE STATUS OF WOMEN IN THE ECONOMICS PROFESSION

Carolyn Shaw Bell, *Chair*
Walter Adams
Francine Blau
Martha Blaxall
Kenneth Boulding
Collette Moser
Barbara Reagan, **Chair*
Myra Strober
Phyllis Wallace
*Florence Weiss

NEW YORK LOCAL ARRANGEMENTS COMMITTEE

Edward Campbell
M. Louise Curley
Louis Fier
Donald R. Gilmore
John Griffin
Irwin Kellner
Seymour M. Kwerel
William E. Kingsley
Martin Lapinsky
Edward Lowry
Robert E. Mack
Edward Marcus
Mildred Marcus
John R. Meyer
Marie Morse
Jean Namias
Robert Ortner
David Pritchett
Philip Watterson
Edward Wolf
Donald Woolley

---

* Newly appointed in 1974.

Council and Other Representatives

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

Stephen Goldfeld (1976)

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SLAVIC STUDIES

Holland Hunter (1973)
Janet Chapman (1976)

AMERICAN COUNCIL OF LEARNED SOCIETIES

Douglass C. North (1974)

FEDERAL STATISTICS USERS' CONFERENCE

John W. Kendrick (1974)

INTERNATIONAL ECONOMIC ASSOCIATION

Fritz Machlup (1975)
John T. Dunlop (1973)
*Abram Bergson (1978)

INTERSOCIETY COMMITTEE ON TRANSPORTATION

Charles A. Taff (1974)

JOURNAL OF RESEARCH ON CONSUMER BEHAVIOR

Mary Jean Bowman (resigned, January 1, 1974)
*Kelvin J. Lancaster (1975)

NATIONAL ARCHIVES ADVISORY COUNCIL—GENERAL SERVICES ADMINISTRATION

Robert Gallman (1975)

NATIONAL BUREAU OF ECONOMIC RESEARCH

Willard L. Thorp (1975)

NATIONAL COUNCIL FOR ACCREDITATION OF TEACHER EDUCATION

Phillip Saunders (1974)

NATIONAL RESEARCH COUNCIL—DIVISION OF BEHAVIORAL SCIENCES

Robert Eisner (1973)
Harold W. Watts (1974)
Barbara Bergmann (1975)

SOCIAL SCIENCE RESEARCH COUNCIL

Albert Rees (1973)
James N. Morgan (1974)
William J. Baumol (1975)
Lawrence R. Klein (1976)

Representatives of the Association on Various Occasions—1973

INAUGURATIONS

Gerson D. Cohen, Jewish Theological Seminary of America
Chaim Ginsberg
David P. Gardner, University of Utah
Allen M. Sievers

RENDIGS FELS, *Secretary*

---

* Newly appointed in 1974.

REPORT OF THE TREASURER FOR THE YEAR ENDING DECEMBER 31, 1973

On December 27, 1973, the Executive Committee approved a budget for the year ending December 31, 1974, with a small anticipated deficit. The budget is shown in Table 1 in comparison with the actual results for 1972, the budget for 1973, and the projections for 1973 available to the Executive Committee when it approved the budget. (The actual results for 1973 are shown below in the financial statements following the Auditors' Report.) The budget includes $70,000 for a new *Handbook* in addition to the reserve of $80,000 accumulated in previous years.

Normal procedures call for a revision of the 1974 budget by the Executive Committee at its March meeting, too late for this report. The revised budget may have a larger deficit in view of the decision taken at the meeting of December 27, 1973, for open listing of job vacancies. This service may incur startup costs not covered by payments for it received in the current year.

Dues and subscriptions were raised 5 percent effective January 1, 1974. In view of the expected deficit this year and the continued high rate of inflation, another increase will probably be needed in 1975.

RENDIGS FELS, *Treasurer*

(Table 1 appears on page 464)

TABLE 1—1974 BUDGET OF THE AMERICAN ECONOMIC ASSOCIATION

Compared to Income Statement for 1972, Budget for 1973, and Income Statement for 1973 Projected from Actual Results for First Nine Months

($ thousands)

	1972 Actual	1973 Budget	1973 Projection	1974 Budget[a]
REVENUES FROM DUES AND ACTIVITIES				
Membership dues and subscriptions	485	500	499	530[b]
Advertising	60	60	79	80
Sale of copies, republications, handbooks	22	20	16	15
Sale of mailing list	16	14	19	25
Annual meeting	13	13	7	15
Other	14	13	11	14
Subtotal	610	620	631	679
INVESTMENT GAINS (LOSSES)	27	19	19	19
TOTAL REVENUE	637	639	650	698
EXPENSES				
Publications				
American Economic Review	160	165	161	178
Journal of Econ. Lit. & Index	203	220	225	235
Papers and Proceedings	39	47	45	48
Subtotal	402	432	431	461
Operating and Administrative				
Committees	6	10	10	15[c]
Salaries and fringe benefits	81	88	78	96
Rent	8	8	8	8
Other	49	50	55	55
Subtotal	144	156	151	174
Total before special items	546	588	582	635
Special Items				
Handbook, reserve for	40	40	40	70
Computerization	—	10	—	—
Committee on Hiring Practices	—	6	6	—
Visiting Scholars Program	—	—	—	3
TOTAL EXPENSES	586	644	628	708
REVENUES OVER (UNDER) EXPENSES	51	(5)	22	(10)

[a] As approved by the Executive Committee on December 27, 1973.

[b] Includes allowance for 5 percent increase in dues and subscriptions effective January 1, 1974.

[c] Includes $5,000 for the Committee on the Status of Women in the Economics Profession.

REPORT OF THE FINANCE COMMITTEE*

The accompanying inventory lists the securities held by the American Economic Association as of December 31, 1973, with the costs and market values as of that date. The total market value of the securities portfolio at year end was $598,958. After making adjustments for cash additions and withdrawals, we estimate that the Association's portfolio (not including the income generated by the account) declined in value by 17.5 percent during 1973. As a basis of comparison, the Dow Jones Industrial Average was down approximately 16.7 percent during the same period. While the Dow Jones Industrial average was down in 1973, its movement does not really reflect the extent of the decline for the market as a whole. In order to place 1973 in perspective, it may be helpful to know that an unweighted average of *all* the stocks on the New York Stock Exchange fell 37.6 percent in 1973. On a six-year comparison, which includes both up and down market cycles, the account has appreciated approximately 20.3 percent, while the Dow Jones Industrial Average has declined 6.1 percent and the unweighted average of New York Stock Exchange stocks has dropped 51 percent.

It should be remembered that the $598,958 figure referred to above includes a Special Grant that was made by the Ford Foundation in January of 1969 and subsequently commingled with the Association's account. As of December 31, 1973, the Association's portion of the aggregate account was $418,911 or 70 percent, and the Special Grant represented the remaining $180,047 or 30 percent of the total.

During the past twelve months, a number of changes were made in the equity portion of the account. This involved the sale of 500 Continental Can, 300 Disney Productions, 500 Times Mirror, 800 Ridder Publications, 500 Western Bancorporation, 300 Squibb, 50 Minnesota Mining, 200 Utah International, 480 First National Bank of Maryland, 300 First Bank Systems, 600 Fort Howard Paper, and 1,000 General Cinema; and new purchases of 700 Newmont Mining, 900 Alexander and Alexander, 800 Manufacturers Hanover Corporation, 700 Columbia Broadcasting, 2,200 Rank Organisation, and 300 Chase Manhattan Corporation.

The year just past was an unusual one in that it contained a particularly large number of unexpected events. Among these were a devaluation of the dollar, a worldwide food shortage, an 8½ percent rate of inflation in consumer prices, war in the Middle East, an energy crisis, and a

TABLE 1—INVENTORY SUMMARY AS OF DECEMBER 31, 1973

	Value	Percent	Estimated Income
Cash and Short-Term Securities	$ 16,935	2.8	$ 2,790
Long-Term Bonds and Preferred Stocks	0	0.0	0
Limited Risk Convertibles	0	0.0	0
Common Stocks and Other Convertibles	582,023	97.2	14,641
Total	598,958	100.0	17,431

---

* The Report of the Finance Committee is informational and is not an audited financial statement. Consequently, there may be some discrepancies between figures in the Report of the Finance Committee and the Auditors' Report which follows.

domestic upheaval, including the resignation of the Vice-President. As we move into 1974, several major uncertainties—including the availability and price of oil, the rate of inflation, the possibility of worldwide recession, and the unsettled domestic political situation—continue to exert pressure on equity prices. However, the managers of the investment fund believe that common stocks, which are down very sharply from peak levels and selling at their lowest relationship to earnings, cash flow, and book value in recent memory, already seem to be discounting an unfavorable resolution of most of the present uncertainties; and that any positive developments, such as a decline in short-term interest rates or the resumption of oil shipments, should provide a real lift for the market. Our overall appraisal of these considerations leads to a reaffirmation of the Finance Committee's basic policy of maintaining a meaningful exposure to equities at this time. Throughout the year, we will continue to reevaluate our investment strategy; and, should a modification in the portfolio's structure appear prudent, proper action will be taken.

BERYL W. SPRINKEL, *Chairman*

TABLE 2—INVENTORY AND APPRAISAL AS OF DECEMBER 31, 1973

	Amount	Price	Value	Unit Cost	Total Cost	Est. Income
CASH AND SHORT-TERM SECURITIES (2.8 percent)						
Cash			($14,065)		($14,065)	
Corporates and Others						
General Finance Demand Notes	31,000	$100	31,000	$100	31,000[a]	$2,790
TOTAL CASH AND FIXED INCOME SECURITIES			16,935		16,935	2,790
COMMON STOCKS AND OTHER CONVERTIBLES (97.2 percent)						
Utilities (5.3 percent)						
Central and Southwest	1,000	17	17,000	8	8,316[a]	1,080
Houston Lighting and Power	500	28	14,000	22	10,942[a]	700
			31,000		19,258	1,780
Banks (15.1 percent)						
Chase Manhattan Corp.	300	56	16,800	56	16,679	600
First Bank System	500	58	29,000	16	7,931	700
Hawaii Bancorporation	1,000	13	13,000	16	16,375	650
Manufacturers Hanover Corp.	800	36	28,800	37	29,763[a]	1,248
			87,600		70,748	3,198
Other Financial (5.5 percent)						
Alexander and Alexander	900	19	17,100	19	16,725[a]	441
ERC Corp.	400	37	14,800	41	16,450	216
			31,900		33,175	657
Papers (2.9 percent)						
Hoerner Waldorf	500	34	17,000	24	12,107	560
Mining and Metals (7.2 percent)						
Newmont Mining	700	32	22,400	27	19,033[a]	980
Utah International Inc.	400	49	19,600	25	10,088	208
			42,000		29,121	1,188

TABLE 2—(*Continued*)

	Amount	Price	Value	Unit Cost	Total Cost	Est. Income
Oil and Gas (15.2 percent)						
Husky Oil CV 6.25 01/15/97	15,000	110	16,500	105	15,750	938
Atlantic Richfield	200	109	21,800	48	9,684	400
Continental Oil	500	55	27,500	33	16,448	800
Kewanee Oil	1,200	19	22,800	17	20,178	576
			88,600		62,060	2,714
Drugs and Medical (2.8 percent)						
Merck	200	81	16,200	45	9,025	280
Electrical Products (2.1 percent)						
Westinghouse Electric	500	25	12,500	32	16,001[a]	486
Computers (8.0 percent)						
International Business Machines	188	247	46,437	112	20,995[a]	842
Other Office Equipment (10.1 percent)						
Moore Corp.	300	55	16,500	37	10,950	240
Rank Organisation	2,200	8	17,600	10	22,513[a]	227
Xerox	200	123	24,600	64	12,814[a]	200
			58,700		46,277	667
Miscellaneous (25.8 percent)						
Columbia Broadcasting	700	26	18,200	30	20,973[a]	1,022
Disney (Walt) Productions	408	47	19,176	3	1,079	49
Eastman Kodak	200	116	23,200	135	26,906[a]	382
Mapco	500	51	25,500	35	17,710	270
Minnesota Mining and Mfg.	300	78	23,400	81	24,448[a]	330
SR&F Cap Opportunities Fund	4,320	9	40,610	12	51,714[a]	216
			150,086		142,830	2,269
TOTAL COMMON STOCKS (100 percent)			582,023		461,597	14,641
TOTAL SECURITIES AND CASH			598,958		478,532	17,431

[a] More than one cost basis.

AUDITORS' REPORT

To the Executive Committee of
The American Economic Association:

We have examined the statements of assets and liabilities of THE AMERICAN ECONOMIC ASSOCIATION (a District of Columbia corporation, not for profit) as of December 31, 1973 and 1972, and the related statements of revenues and expenses, changes in general and restricted fund balances and changes in assets and liabilities for the years then ended. Our examination was made in accordance with generally accepted auditing standards, and accordingly included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the accompanying statements present fairly the assets and liabilities of The American Economic Association as of December 31, 1973 and 1972, and its revenues and expenses, changes in fund balances and the changes in its assets and liabilities for the years then ended, in conformity with generally accepted accounting principles consistently applied during the periods.

ARTHUR ANDERSEN & CO.

Memphis, Tennessee,
March 5, 1974.

THE AMERICAN ECONOMIC ASSOCIATION

STATEMENTS OF ASSETS AND LIABILITIES—DECEMBER 31, 1973 AND 1972

ASSETS	1973	1972	LIABILITIES AND FUND BALANCES	1973	1972
CASH, including $145,000 certificate of deposit in 1973	$ 208,945	$117,547	ACCOUNTS PAYABLE AND ACCRUED LIABILITIES	$ 101,445	$ 96,528
			DEFERRED INCOME (Note 1):		
INVESTMENTS, at market (cost of $665,688 and $565,938) (Notes 1 and 2)	776,023	814,774	Life membership dues	49,794	43,264
			Prepaid membership dues	205,209	129,136
			Prepaid subscriptions	133,802	88,963
RECEIVABLES, less allowance for doubtful accounts of $2,000	24,496	25,610	Total deferred income	388,805	261,363
PREPAID EXPENSES	2,043	1,066			
			ACCRUED PUBLICATION EXPENSES (Note 1)	80,000	40,000
OFFICE FURNITURE AND EQUIPMENT, at cost, less accumulated depreciation of $3,423 and $2,664	6,906	7,013	FUND BALANCES:		
			Restricted	313,502	301,474
			General	226,610	203,344
DEFERRED CHARGES (CREDITS)—unrecognized change in market value of investments (Notes 1 and 3)	91,949	(63,301)	Total fund balances	540,112	504,818
	$1,110,362	$902,709		$1,110,362	$902,709

The accompanying notes to financial statements are an integral part of these statements.

THE AMERICAN ECONOMIC ASSOCIATION
STATEMENTS OF REVENUES AND EXPENSES
FOR THE YEARS ENDED DECEMBER 31, 1973 AND 1972

	1973	1972
REVENUES FROM DUES AND ACTIVITIES:		
Membership dues	$311,961	$311,766
Subscriptions	194,526	173,313
Advertising	72,930	59,540
Sale of copies, republications and handbooks	17,765	22,282
Sale of mailing list	20,112	16,095
Annual meeting	7,491	13,383
Sundry	11,289	14,360
	636,074	610,739
INVESTMENT GAINS (Note 2)	1,135	26,732
Total revenues	637,209	637,471
PUBLICATION EXPENSES:		
American Economic Review	172,161	160,592
Journal of Economic Literature	221,450	203,525
Papers and Proceedings	45,042	38,555
Provision for Directory publication (Note 1)	40,000	40,000
	478,653	442,672
OPERATING AND ADMINISTRATIVE EXPENSES:		
Committee	16,550	5,780
Annual meeting	12,825	6,661
General and administrative:		
Salaries	86,200	80,936
Rent	8,005	7,896
Other (Exhibit I)	49,321	40,100
Interest and miscellaneous	1,039	2,371
	173,940	143,744
Total expenses	652,593	586,416
REVENUES IN EXCESS OF (LESS THAN) EXPENSES	$(15,384)	$ 51,055

The accompanying notes to financial statements and Exhibit I are an integral part of these statements.

THE AMERICAN ECONOMIC ASSOCIATION
STATEMENT OF CHANGES IN GENERAL FUND BALANCE
FOR THE YEARS ENDED DECEMBER 31, 1973 AND 1972

	Total	Operations	Market Value Adjustments
Balance at January 1, 1972	$142,121	$ 26,513	$115,608
Add—			
Market value adjustments arising from inflation	10,168	—	10,168
Revenues in excess of expenses	51,055	51,055	—
Balance at December 31, 1972	203,344	77,568	125,776
Add—market value adjustments arising from inflation	38,650	—	38,650
Deduct—expenses in excess of revenues	(15,384)	(15,384)	—
Balance at December 31, 1973	$226,610	$ 62,184	$164,426

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION
STATEMENT OF CHANGES IN RESTRICTED FUND BALANCES
FOR THE YEAR ENDED DECEMBER 31, 1973

	Balance at January 1	Receipts	Disbursements	Allocation of Investment Gains (Note 4)	Balance at December 31
The Ford Foundation grants for—					
Economics Institute's orientation program for foreign graduate students of economics	$257,342	$24,594	$(16,230)	$14,795	$280,501
Committee on the Status of Women in the Economics Profession	15,571	—	(14,551)	—	1,020
The Alfred P. Sloan Foundation and Chase Manhattan Bank grants for increase of educational opportunities for minority students in economics	25,000	—	(1,097)	—	23,903
Funds reserved by the Association for publication of revised editions of *Graduate Study in Economics*, a guide originally published with funds from a Ford Foundation grant	3,240	2,792	—	—	6,032
The Asia Foundation grant for Asian economist membership dues to The American Economic Association and related travel expenses	10	4,149	(2,524)	—	1,635
Sundry	311	100	—	—	411
	$301,474	$31,635	$(34,402)	$14,795	$313,502

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION
STATEMENT OF CHANGES IN RESTRICTED FUND BALANCES
FOR THE YEAR ENDED DECEMBER 31, 1972

	Balance at January 1	Receipts	Disbursements	Allocation of Investment Gains (Note 4)	Balance at December 31
The Ford Foundation grants for—					
Translation of foreign economic publications and survey of foreign economic research	$ 1,584	$ 431	$ (2,015)	$ —	$ —
Economics Institute's orientation program for foreign graduate students of economics	198,922	42,338	—	16,082	257,342
Committee on the Status of Women in the Economics Profession	—	25,000	(9,429)	—	15,571
The Alfred P. Sloan Foundation and Chase Manhattan Bank grants for increase of educational opportunities for minority students in economics	—	25,000	—	—	25,000
Funds reserved by the Association for publication of revised editions of *Graduate Study in Economics*, a guide originally published with funds from a Ford Foundation grant	3,240	—	—	—	3,240
The Asia Foundation grant for Asian economist membership dues to The American Economic Association and related travel expenses	1,163	—	(1,153)	—	10
The National Science Foundation grant for visiting scientist program	190	8,445	(8,635)	—	—
Sundry	246	115	(50)	—	311
	$205,345	$101,329	$(21,282)	$16,082	$301,474

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION
STATEMENTS OF CHANGES IN ASSETS AND LIABILITIES
FOR THE YEARS ENDED DECEMBER 31, 1973 AND 1972

	1973	1972
SOURCE (USE) OF FUNDS:		
Revenues in excess of (less than) expenses	$(15,384)	$ 51,055
Add—		
Realized gains from sales of investments	1,619	8,051
Noncash charges (credits):		
Market value adjustments (Note 1)	18,821	(18,173)
Depreciation	800	746
Accrued publication expenses (Note 1)	40,000	40,000
Funds provided by operations	45,856	81,679
(Increase) decrease in—		
Investments:		
Proceeds from sales	394,135	412,586
Purchases, at cost	(483,885)	(602,730)
Receivables and prepaid expenses	137	10,837
Office furniture and equipment	(693)	(285)
Increase (decrease) in—		
Note payable to bank	—	(25,000)
Accounts payable and accrued liabilities	4,917	16,824
Deferred income	127,442	33,061
Restricted funds	3,489	90,938
Increase in cash	$ 91,398	$ 17,910

The accompanying notes to financial statements are an integral part of these statements.

THE AMERICAN ECONOMIC ASSOCIATION
NOTES TO FINANCIAL STATEMENTS

(1) Significant Accounting Policies:

Investments—

The Association accounts for its investments on a market value basis. Under the method used by the Association to value these investments, the change in market value of corporate stocks during the year, after adjusting for an inflation factor, is recognized in income over a three-year period. The change in market value of corporate bonds and obligations, after adjusting for this inflation factor, is reflected in income currently. The portion of changes in market value of investments attributable to inflation is allocated to the general and restricted fund balances as appropriate.

Accrued Publication Expenses—

Approximately every three to five years, The American Economic Association publishes a Directory which lists, among other things, the names and addresses of its membership. This Directory is distributed at no additional cost to the membership in the year of publication. In order to more properly match the publishing cost of this Directory with revenue from membership dues, the Association is providing, annually, a reserve for estimated costs to be incurred when the Directory is again published. Actual costs, when incurred, will be charged to this reserve.

Deferred Income—

Revenue from membership dues and subscriptions to the various periodicals of the Association are deferred when received, with these amounts then recognized as income quarterly as publications are mailed to the members and subscribers.

Life membership dues are also deferred when received, with subsequent recognition as income occurring over the estimated average life of these members.

(2) Investments and Investment Income:

The following is a summary of investments held by the Association at December 31:

	1973		1972	
	Cost	Market	Cost	Market
Corporate bonds and obligations	$209,750	$210,500	$121,750	$122,650
Corporate stocks	445,938	565,523	444,188	692,124
	$655,688	$776,023	$565,938	$814,774

Investment gains recognized in income for the years ended December 31, were as follows:

	1973	1972
Corporate bonds and obligations—		
Interest	$13,270	$10,566
Decline in market value	(6,649)	(853)
	6,621	9,713
Corporate stocks—		
Cash dividends	12,943	8,884
Increase (decline) in market value recognized (Note 3)	(20,277)	19,026
	(7,334)	27,910
Other-investment (gains) losses allocated to a restricted fund (Note 4)	1,848	(10,891)
Investment gains included in income	$ 1,135	$26,732

(3) Deferred (Charges) Credits:

As described more fully in Note (1), the Association recognizes in income over a three-year period changes in the market value of its corporate stocks. The following summarizes the years in which market value changes of stocks occurred that affect 1972 and 1973 revenues, and the amount of these market value increases (declines) that have been or will be recognized in income:

Year of Market Value Change	Recognized In Income In		Deferred Until		Balance Deferred as of December 31	
	1972	1973	1974	1975	1973	1972
1970	$(19,206)	$ —	$ —	$ —	$ —	$ —
1971	13,163	13,163	—	—	—	13,163
1972	25,069	25,069	25,069	—	25,069	50,138
1973	—	(58,509)	(58,509)	(58,509)	(117,018)	—
	$ 19,026	$(20,277)	$(33,440)	$(58,509)	$(91,949)	$63,301

Included in the above totals are amounts which have been or will be allocated to a restricted fund. The amounts allocated are based on the percentage of the Association's total stock portfolio owned by this restricted fund, times the above market value changes.

(4) Restricted Fund:

In 1968, the Association entered into an agreement with the University of Colorado relating to the Ford Foundation grant for the Economics Institute which provides, among other things, that the Association invest a portion of the funds received and allocate any income and market value adjustments therefrom to the restricted fund. In accordance with this agreement, the following net investment gains (losses) and other adjustments were allocated to the restricted fund:

	1973	1972
Investment gains (losses) (Note 2)	$(1,848)	$10,891
Market value adjustments arising from inflation	16,643	5,191
	$14,795	$16,082

(5) The Association:

The American Economic Association is an educational organization, substantially exempt from income tax under section 501 (c) (3) of the U.S. Internal Revenue Code. As required by Section 511 (a) of this Code, the Association does provide for Federal income taxes on certain revenues which are not substantially related to its tax exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists.

The Association has been determined to be an organization which is not a private foundation.

EXHIBIT I—THE AMERICAN ECONOMIC ASSOCIATION
STATEMENTS OF OTHER GENERAL AND ADMINISTRATIVE EXPENSES
FOR THE YEARS ENDED DECEMBER 31, 1973 AND 1972

	1973	1972
Mailing list file maintenance and periodic mailing expenses	$14,205	$14,410
Accounting and legal	7,300	5,537
Office supplies	5,507	5,507
Postage	5,383	4,857
Dues and subscriptions	3,325	2,725
Telephone	3,063	2,762
Provision for Federal income taxes (Note 5)	3,000	—
Investment counsel and custodian fees	2,865	2,517
President and president-elect expenses	2,334	678
Travel and entertainment	1,084	646
Depreciation	800	746
Uncollectible receivables	262	—
Currency exchange charges	193	183
Expenses allocated to a restricted fund	—	(468)
	$49,321	$40,100

REPORT OF THE MANAGING EDITOR AMERICAN ECONOMIC REVIEW

The number of manuscripts submitted in 1973 was 758, an increase over the number sent in 1972. We printed 111 papers in all, which is fewer than last year. These statistics are shown in Table 1, together with a record of the past twenty years.

Manuscripts Submitted, Accepted, and Published

After two years of decline from 879 in 1970, the number of papers has started to increase. The present methods of screening and refereeing papers are sufficiently elastic to allow for increased numbers, and apparently authors have become accustomed and are willing to accept the time required by the manuscript review process. These delays are inevitable, and I am grateful for the patience which authors display while waiting for critical evaluation and decision on their papers. Some of the delays are in my office and others occur with the referees. We follow the procedure of withdrawing a manuscript from a referee if he fails to respond within six weeks. We then send the paper to a new referee. This cuts down the extent to which a paper is hung up because of delay at a single referee. Long delays occur for a different reason. In rare cases I will have to try three or even four referees before I find one who has the time and interest to read and criticize a paper. Thus refereeing is a chief source of delay, but not for the reason that many think.

TABLE 1—MANUSCRIPTS SUBMITTED AND PUBLISHED, 1953–73

Year	Submitted	Published	Ratio of Published to Submitted
1953	234	51	.22
1954	231	47	.20
1955	245	41	.17
1956	242	48	.20
1957	215	40	.19
1958	242	46	.19
1959	279	48	.17
1960	276	46	.17
1961	305	47	.15
1962	273	46	.17
1963	329	46	.14
1964	431	67	.16
1965	420	59	.14
1966	451	62	.14
1967	534	94	.18
1968	637	93	.15
1969	758	121	.16
1970	879	120	.14
1971	813	115	.14
1972	714	143	.20
1973	758	111	.15

The backlog of accepted papers has increased slightly. As of November 30 there are seventy-nine accepted papers which have not yet been published. Thirty will appear in the March 1974 issue, twenty-six in June 1974, and the remaining twenty-three will appear September and December 1974.

We printed a larger number of main articles this year than last, and their average size was slightly longer. For this reason fewer short papers were printed this year, and the total number (main plus short papers) printed declined. I am concerned by this development and will seek to reverse it. One method will be to reduce the size of accepted papers. Table 2 summarizes the contents of the four regular issues of the *Review* in 1973, comparing it with the previous year. It can be seen that we printed 1,146 pages in all, because of a very large December issue. I don't expect a December bulge to be a regular feature of the *Review*. A number of factors contributed to it which will not recur. For example, the dissertation section of sixty-seven pages was enlarged by the inclusion

TABLE 2—SUMMARY OF CONTENTS, 1972 AND 1973

	1972		1973	
	Number	Pages	Number	Pages
Articles	47	588	54	747
Shorter papers, including notes, comments, and replies	96	396	57	273
Special Articles	2	9	2	15
Dissertations		43		67
Announcements and notes		45		44
Total	145	1,081	113	1,146

of an unusual number of 1972 listings that had not been reported last year. The bulge also occurred because of a last minute decision to include a number of papers in December rather than run them in 1974.

Subject Matter of Submitted and Printed Manuscripts

The subject matter distribution of submitted and published manuscripts is shown in Table 3. The most popular fields remain microeconomics, macro and monetary theory, international economics, labor economics, and welfare economics. The distribution of published papers reflects the distribution of submitted papers. Each year when this report is delivered to the annual meeting, the question inevitably arises why the distribution of published articles is what it is. Why, for example, does the *Review* print so many articles employing neoclassical techniques, or why so many articles using mathematics? Why, for example, is the work of the radical economists largely ignored? Why doesn't the editor commission articles in these

TABLE 3—SUBJECT MATTER DISTRIBUTION OF SUBMITTED AND PUBLISHED MANUSCRIPTS IN 1973

	Submitted	Published
General Economics and General Equilibrium Theory	9	2
Microeconomic Theory	108	7
Macroeconomic Theory	49	5
Welfare Theory	51	13
Economic History, History of Thought, Methodology	12	4
Economic Systems	17	—
Economic Growth, Development, Planning, Fluctuations	44	8
Economic Statistics	37	5
Monetary and Financial Theory and Institutions	61	9
Fiscal Policy and Public Finance	25	2
International Economics	81	17
Administration, Business Finance	37	4
Industrial Organization	44	7
Agriculture, Natural Resources	34	3
Manpower, Labor, Population	96	15
Welfare Programs, Consumer Economics, Urban and Regional Economics	52	10
Other	1	—
Total	758	111

underrepresented fields? For if people aren't working in these areas, they should be encouraged to do so. I discussed this issue in my report last year, and I should like to mention it again. We are by and large bound to print what is sent in, and we try to cull out the best there is by refereeing. There is no way to guarantee the quality of commissioned articles, so that in the main what the *Review* prints reflects what the profession is doing. There are nevertheless no biases against papers in underrepresented fields. I also spelled this position out in a statement of editorial policy printed in the September 1972 *Review* and again in the May 1973 *Papers and Proceedings* issue.

Anonymous Refereeing

At the December 1972 annual meeting in Toronto, a member moved the adoption of a resolution calling for anonymous refereeing of manuscripts. Upon addressing the meeting in favor of the motion, its proponent argued that anonymity would protect the relatively unknown author. He put forward the argument that referees tend to favor and pay attention to the works of better known writers and give less care when reading the lesser known. There was no way of assessing the truth of these arguments on an a priori basis. In replying I requested that the motion be altered to allow an experiment, under which half of the submitted papers would be refereed anonymously. The other half would be refereed as before with the author's name remaining on the paper. I promised to report on the results of this procedure at the December 1973 annual meeting.

Because of time delays in preparing reports, the experience discussed below covers only the 554 papers submitted between January 1 and September 28, 1973. Of these 554, a decision had been reached on 311. The remaining 243 were still in the hands of referees or screeners, or awaiting disposition. The results are as follows:

	Author's Name on Paper	Anonymous	Total
Submitted	277	277	554
Action Taken	160	151	311
Rejected	144	133	277
Accepted	4	4	8
Accepted subject to revision	12	14	26
Successfully passed screening	61	72	133

It is clear that there is little difference in treatment accorded the two types of papers. If any differentiation occurred it was in favor of anonymous authors. Note that 61 of 160 known author papers survived screening, while 72 of 151 anonymous papers survived. As a piece of qualitative information, all of the referees who commented on the practice of anonymity were favorable. This included cases where the referee recognized the name of the author. Recognition occurred in the case of 20 papers, and did so either because the referee had, in addition, received the paper directly from the author, or because the author repeatedly cited his own work in the text. The latter practice is a loophole in the process of anonymous refereeing. The referees were favorably disposed toward anonymous refereeing. They felt it was inherently fair and would dispel suspicion of differential treatment. Some acknowledge that they worked harder on anonymous papers, where they might have succumbed to the temptation to accept and not question obscure argumentation from better known authors. In no case did a referee complain that ability to judge a paper was hindered by the absence of the author's name.

My conclusion is that anonymous ref-

TABLE 4—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING IN 1973

	Number of Copies Printed	Pages		Cost		
		Net	Gross	Issue[a]	Reprints	Total
March.......	26,000	246	320	$ 25,527.15	$ 310.31	$25,837.46
June.........	27,000	262	312	26,737.87	466.59	27,204.46
September....	26,000	278	320	28,587.49	282.79	28,870.28
December....	27,000	360	432	35,552.26	408.65	35,960.91
Total........	106,000	1,146	1,384	$116,404.77	$1,468.34	$117,873.11

[a] Includes allocated cost of preparing mailing list.

ereeing can work. I still do not have a full year of experience on which to make this judgment, and it is possible that unforeseen difficulties may occur with the processing of anonymous papers. For example, the determined (and well-known) author can always confound the system through repeated self-referencing. But the evidence to date indicates that it is acceptable to referees.

Expenses—Printing and Mailing

The 1973 costs of printing and mailing the *Review* are shown in Table 4 and shown for other years in Table 5. The net number of pages includes only those pages listed in our table of contents. Advertising pages then make up the difference between gross and net. The 1973 printing expenses will run close to the estimate. The printing contract is again up for renegotiation and I expect higher costs next year. Paper costs have already risen, and our 1973 bills now contain a paper surcharge. To some degree, the increased printing costs for next year can be offset by more economical methods of preparing the *Review* for publication. For example, printing a front portrait on regular stock rather than on enamel paper can save over 2 percent of the cost of an entire issue. The forecast of next year's printing costs shown in Table 5 will remain a guess until our negotiations with the printer move further along.

Office Expenses

Office expenses rose in 1973 because of higher salaries and increased supply costs. The 1973 figure shown in Table 5 is still an estimate, as all charges are not yet in. I

TABLE 5—ACTUAL AND BUDGETED EXPENDITURES, 1966–1974

	Printing and Mailing	Payments to Contributors	Office Expenses	Total
1966	$ 74,277	$1,873	$24,844	$100,996
1967	82,840	2,255	27,114	112,209
1968	92,948	2,140	29,155	124,243
1969	97,183	600	26,765	124,548
1970	111,227		36,336	147,564
1971	120,120		43,524	163,644
1972	107,196		44,473	151,669
1973 (budget)	115,000		49,283	164,283
1973 (actual)	115,372[a]		49,121	164,493
1974 (budget)	121,518		53,645	175,163

[a] Estimate.

expect office expenses to rise again in 1974 because of higher salaries and fringe benefit costs.

Board of Editors

Six members of the Board of Editors complete their terms at this time: Barbara Bergmann, Phillip Cagan, Carl Christ, Daniel McFadden, Herbert Mohring, and Edmund Phelps. I am deeply indebted to them for their hard work, high critical standards, and cooperation. They have helped to maintain and improve the quality of the *Review*.

I am submitting to the Executive Committee the names of six nominees to serve three-year terms, until December 31, 1976. They are Martin S. Feldstein, Bent Hansen, Jerome Stein, S. C. Tsiang, Finis Welch, and Marina v.N. Whitman.

I should like to express my thanks to the continuing members of the Board of Editors: Bela Balassa, Robert Baldwin, Ann Friedlaender, Stephen Goldfeld, Robert Hall, Jack Hirshleifer, Leonard Rapping, Sherwin Rosen, Thomas Saving, Anna Schwartz, and Joseph Stiglitz. I appreciate their work and assistance.

The following assisted me this year as editorial consultants: Robert Barro, Raveendra Batra, Roger Betancourt, Walter Dolde, Allan Feldman, Stanley Fischer, James Hanson, E. Philip Howrey, John Kennan, John Ledyard, Robert Rohr, Thomas Sargent, and Daniel Wisecarver.

I have received valuable proofreading and mathematical assistance from: Clemens F. J. Boonekamp, Hajime Hori, Robert Jones, Hiroshi Ono, Itchak Dov Paris, Rama Ramachandran, Orlando Roncesvalles, and John Seater. Finally, I am grateful to Wilma St. John for her fine work as assistant editor.

In addition to the members of the Board and the editorial consultants, I have sought and received the assistance of a large number of economists during the course of the year. I wish to thank them for their cooperation and application of high standards in reading and evaluating manuscripts. They have eased the heavy work load that would have otherwise fallen on the Board of Editors. The following have assisted in this way:

I. Adelman
B. Aghevli
D. Aigner
N. Aitken
A. Alchian
P. Allen
J. Anderson
G. C. Archibald
K. J. Arrow
O. Ashenfelter
E. Bailey
M. Balch
R. Barlow
D. P. Baron
Y. Barzel
W. J. Baumol
R. Beals
G. S. Becker
M. Beckmann
J. R. Behrman
F. Bell
G. Benston
T. J. Bertrand
S. M. Besen
G. O. Bierwag
M. Blume
R. G. Bodkin
T. E. Borcherding
M. Boskin
R. Bower
S. Bowles
J. Brada
D. Bradford
W. H. Branson
R. A. Brecher
F. P. R. Brechling
M. Brennan
W. Brock
M. Bronfenbrenner
G. Brown
M. Brown
E. K. Browning
H. J. Bruton
J. M. Buchanan
O. Burt
G. Cain
J. A. Carlson
F. Casas
D. Cass
M. Chacholiades
S. Cheung
G. Childs
B. R. Chiswick
G. C. Chow
L. Christensen
C. Cicchetti
C. S. Clark
P. B. Clark
R. Clower
R. Coen
W. Comanor
M. Connolly
J. P. Cooper
P. Cootner
T. J. Courchene
J. Craft
J. C. Cragg
R. G. Cummings
J. Cunnyngham
G. Daly
M. R. Darby
R. d'Arge
K. Davis
O. Davis
R. H. Day
F. de Leeuw
G. De Menil
H. Demsetz

E. Denison
M. E. DePrano
M. De Tray
D. Dewey
P. A. Diamond
R. K. Diwan
P. Doeringer
R. Dorfman
G. Douglas
W. C. Dunkelberg
G. Eads
R. Eckaus
R. Ehrenberg
I. Ehrlich
B. Ellickson
J. W. Elliott
K. Elzinga
J. Enzler
W. Ethier
R. Evenson
E. Fama
E. L. Feige
M. Feldstein
W. Fellner
S. Fischer
A. Fisher
F. Flatters
B. Fleisher
M. Frankel
R. Freeman
J. Fried
G. Fromm
E. Furubotn
M. Galatin
F. Gehrels
G. Ghez
W. E. Gibson
C. Goetz
S. Goldman
M. J. Gordon
R. J. Gordon
J. P. Gould
H. Grabowski
E. Gramlich
E. Greenberg
M. Greenhut
R. Grieson
J. Griffin
R. Gronau
H. Grossman
M. Grossman
M. A. Grove
J. G. Gurley
J. M. Guttentag
M. Hadjimichalakis
K. Hamada
R. Hamada
M. Hamburger
D. Hamermesh
B. Hansen
W. L. Hansen
J. Harris
J. C. Hause
T. Havrilesky
G. A. Hay
J. Heckman
W. P. Heller
P. Helmberger
D. W. Henderson
J. Hixson
D. Hodgman
R. Holbrook
C. C. Holt
R. Hooley
H. Hori
T. Horst
G. Horwich
E. S. Howle
F. S. T. Hsiao
G. C. Hufbauer
D. T. Hulett
S. H. Hymans
B. Imel
J. C. Ingram
D. Jaffee
E. James
F. Jen
M. C. Jensen
M. B. Johnson
T. Johnson
R. W. Jones
D. W. Jorgenson
M. Kafoglis
J. F. Kain
H. Kaitz
M. Kamien
E. J. Kane
N. Kaplan
G. G. Kaufman
D. Keesing
M. G. Kelly
M. Keren
R. A. Kessel
J. Kindahl
K. Klappholz
M. Klein
A. K. Klevorick
J. Kmenta
T. Koizumi
T. C. Koopmans
D. T. Kresge
C. G. Krouse
A. Krueger
J. V. Krutilla
R. Kuenne
P. Kumar
H. Kunreuther
M. Kurz
M. Ladenson
A. Laffer
D. Laidler
K. J. Lancaster
W. Landes
L. J. Lau
L. B. Lave
S. Lebergott
C. Lee
T. H. Lee
A. Leibowitz
H. Leland
D. Levhari
H. Levin
J. Levin
R. Lind
C. M. Lindsay
R. Litzenberger
N. Liviatan
T. Lowinger
R. Lucas
I. Lurie
R. A. McCain
J. J. McCall
J. McGowan
L. McKenzie
R. McKinnon
G. S. Maddala
S. P. Magee
J. N. Makin
A. Manne
A. L. Marty
R. T. Masson
T. Mayer
A. H. Meltzer
J. R. Melvin
C. Metcalf
R. A. Meyer
P. Mieszkowski
D. D. Milenkovitch
M. Miller
N. Miller
E. S. Mills
J. Mincer
H. Minsky
E. J. Mishan
M. Montias
J. H. Moore
D. T. Mortensen
B. Motley
D. Mueller
M. Mussa
R. Muth
E. Nadel
K. Nagatani
S. Naya
P. Neher
R. Nelson
M. L. Nerlove
J. Newhouse
P. K. Newman
D. Nichols
J. Niehans
Y. Niho
R. Noll
W. Nordhaus
W. Oakland
W. E. Oates
B. Okner
E. O. Olsen

J. Olson
M. Olson
Y. Onitsuka
D. Orr
D. K. Osborne
H. Pack
R. E. Park
R. Parks
D. Parsons
M. Parvin
A. Pascal
P. Pashigian
M. V. Pauly
S. Pejovich
S. Peltzman
J. Pencavel
R. G. Penner
R. Perlman
G. Perry
W. Peterson
G. Pidot
G. Pierson
T. Pinfold
J. E. Pippenger
C. G. Plourde
R. Pollak
W. Poole
R. D. Portes
F. L. Pryor
F. G. Pyatt
R. E. Quandt
J. Quirk
T. Rader
F. Raines
R. H. Rasche
A. Ray
A. Razin
U. E. Reinhardt
S. Reiter
S. Resnick
M. K. Richter
M. Ringo
J. D. Rodgers
R. Roll
D. Roper
S. Ross
J. Rosse
T. J. Rothenberg
M. Rothschild
S. Rottenberg
J. C. R. Rowley
V. Ruttan
H. E. Ryder
N. B. Ryder
D. Salkever
P. A. Samuelson
A. Sandmo
A. Santomero
R. Saposnik
T. Sargent
M. Sarnat
R. Sato
M. Sattinger
J. Scadding
F. M. Scherer
R. Schmalensee
R. Schuler
T. P. Schultz
C. L. Schultze
D. Schwartzman
G. W. Scully
W. F. Sharpe
J. B. Shepherd
J. Shiskin
C. D. Siebert
W. Silber
E. Silberberg
H. A. Simon
C. A. Sims
L. A. Sjaastad
D. Sjoquist
V. Smith
E. Smolensky
D. Smyth
R. Soligo
L. Solmon
R. M. Solow
H. F. Sonnenschein
J. Stein
A. Steinherr
H. O. Stekler
R. M. Stern
G. Stevens
D. Stewart
J. A. Stockfisch
R. Sutch
A. Takayama
J. E. Tanner
P. J. Taubman
M. Taussig
L. D. Taylor
L. Telser
M. Teubal
H. Theil
R. W. Thomas
J. Thornton
A. W. Throop
L. Thurow
T. N. Tideman
T. H. Tietenberg
P. A. Tinsley
J. Tobin
M. P. Todaro
G. Tolley
R. D. Tollison
W. Travis
S. C. Tsiang
D. Tucker
J. Vanderkamp
D. Vandermeulen
J. Vanek
E. C. H. Veendorp
W. S. Vickery
M. Visscher
G. von Furstenberg
T. J. Wales
F. V. Walker
G. Walton
H. Wan
H. W. Watts
R. Waud
J. Weicher
B. A. Weisbrod
M. L. Weitzman
F. R. Welch
S. H. Wellisz
F. W. Westfield
R. Wichers
J. Wile
J. G. Williamson
O. Williamson
R. J. Willis
S. G. Winter
D. Wisecarver
A. D. Witte
J. G. Witte
B. Wolkowitz
D. A. Worcester
S. Y. Wu
M. E. Yaari
W. P. Yohe
P. A. Yotopoulos
E. Zabel
P. Zarembka
R. Zeckhauser

GEORGE H. BORTS, *Chairman*

REPORT OF THE MANAGING EDITOR
JOURNAL OF ECONOMIC LITERATURE

Again, the annual review presents an opportunity to report on the year's output, to identify its particular problems, and to outline some plans for the future.

Table 1 illustrates our product. Members will note that we have published a larger journal than ever before. However, it is clear that most of the expansion is associated with our listing and indexing of a growing number of journals. A final explanation for the expansion of our number of pages is the somewhat larger number of abstracts of articles which we have chosen to print. We have abstracting agreements with a good many journals, having inherited most of them from our previous incarnation (the *Journal of Economic Abstracts*). We have reduced the size of abstracts from 300 to 100 words and we have, in the interests of equity, asked most editors to submit abstracts (where an abstracting agreement exists) of no more than half the articles they print. In brief, 1973 has been a struggle against the expansion of the journal—a struggle made necessary by limitations of staff, budget, and the (to us, at least) unanticipated high increase in the cost of printing.

During 1973 we published four survey articles, which is our target. We published four other major articles, dividing them equally between reflective essays on subfields and articles about economic literature. The number of communications we had was much greater than ever before; however, they were also much shorter. We annotated approximately the same number of books as in previous years and

TABLE 1—QUANTITATIVE ANALYSIS OF CONTENTS

Journal of Economic Literature, 1969–1973

	1969		1970		1971		1972		1973	
	Number	Pages	Number	Pages	Number	Pages	Number	Pages	Number	Pages
Survey articles	4	111	4	123	1	37	3	78	4	144
Essays on subfields	3	55	4	82	5	105	5	129	2	28
Review articles	4	47	6	39	5	40	5	39	—	—
Articles about economic literature	1	12	—	—	2	32	1	19	2	38
Communications	2	4	7	18	7	24	4	13	10	26
Books annotated	1,203	182	1,145	171	1,279	206	1,209	257	1,214	239
Books reviewed	188	243	177	259	172	249	160	241	175	259
Journal issues listed and indexed	668	101	717	125	835	130	849	140	1,011	185
Number of individual articles							5,387		7,218	
Subject index of journal articles	—	195	—	296	—	251	—	263	—	357
Abstracts of articles	1,139	444	1,181	356	1,426	337	1,415	313	1,906	407
Total pages[a]		1,432		1,539		1,479		1,572		1,742[b]

[a] Includes, in addition to listed pages, classification systems, table of contents, indices, journal subscription information, etc.

[b] Estimate.

we also reviewed about the same number as previously. One lesson we have learned is that group reviews (a single review covering several books) are hard to write and, consequently, hard to commission.

Work has continued on the concurrent project of producing annual indexes. The pipeline has been established and the 1970 manuscript is being shipped to our new printer, the Kingsport Press, Kingsport, Tennessee. One reason for the decision to move to Kingsport from the George Banta Company, Menasha, Wisconsin, is that Kingsport has the computer capability to process the material we need for the quarterly index (as found in the *Journal of Economic Literature*) and the annual indexes (this is the quarterly material supplemented by material from *Festschrifts*, works of collected essays, and selected government documents). Over the past two years in an effort to preserve retrieval possibilities, we have (1) used a computer terminal in our offices to process the classified citations; (2) shipped the product to be processed by the National Bureau of Economic Research on the Yale University computer; (3) received and corrected the error runs; (4) received, in due course, the necessary sorted printouts; (5) prepared them for the printer; (6) finally sent the printouts to be set by the conventional typesetting means at the Menasha plant of the George Banta Company. Our new procedure provides that we will no longer be handling the computer input at our offices; rather, we will send the material in a very simple typescript fashion to be put on the Kingsport computer which, in time, will process the data both for the *Journal of Economic Literature* (the quarterly) and the annual indexes. Kingsport will preserve the tapes for us for any future use.

This is, therefore, a splendid opportunity to thank John Meyer and Orin Hansen of the National Bureau of Economic Research for the extraordinarily generous cooperation they have provided in processing our index material. Their efforts have been costly in their time and certainly reflect an optimum level of cooperation between an outstanding economic research institution (the National Bureau of Economic Research) and one agency (the *Journal of Economic Literature*) of this Association.

I should also renew publicly our thanks to James Williams (now at the University of South Carolina-Columbia) for his continuous willingness to help us with our programming problems. He has been consulted both by the National Bureau of Economic Research and the Kingsport Press. In every instance his willingness to take time (at his own expense) to help resolve knotty problems goes far beyond any call of friendship or professional responsibility. I should also add that the Administration of the University of Pittsburgh (particularly the Chancellor and the Dean of the Faculty of Arts and Sciences) has helped the American Economic Association by the assignment of some Teaching Assistants plus the allocation of space and equipment.

While it has not been the custom to thank individually the clerical staff of the *Journal*, the fact is that our three principal secretaries (Mrs. Lyndis Rankin, Mrs. Peggy Yanchosek, and Mrs. June Harrison) have processed a much larger journal than has heretofore been the case; they have done this not only by increasing their hourly productivity, but, in the case particularly of Mrs. Harrison, by working unpaid extra hours.

Each issue of the *Journal* carries the names of students both in Pittsburgh and elsewhere who have helped us with the editorial processes. We, as the members of the Association are aware, annotate all books sent to us. These books are annotated by students. The Associate Editor trains the annotators and both she (the

Associate Editor) and I check every annotation. It is a time-consuming process. During 1973 we annotated no less than around 1,400 books. In many instances the annotations, as originally given to us, had to be redone because the students involved were still in the training process. The Associate Editor also supervises the classification of material for the quarterly and annual indexes. She trains the classifiers, who are generally second- and third-year graduate students. The students' work is then reviewed by several of my colleagues; these have included Asatoshi Maeshiro, Michael Spiro, and Jerome Wells. Their product (insofar as any questions remain) is then reviewed by the Associate Editor.

At each year's convention questions are raised about how we select books for review and how we select reviewers. By and large we have space to review only approximately 170 to 180 books per year. Generally the process involves an attempt to diversify subject areas (making sure each subfield has some books reviewed) and to identify books for review because they are highly original and carry important contributions to our knowledge of data, theory, and policy. We also try to influence our selection by choosing books written by relatively young economists. We are not always successful in the latter effort, but it is worth noting that a conscious effort is made in this direction. When it comes to choosing a reviewer, we, like all editors, prefer to diversify our reviewers, and again we prefer to choose young economists if that is feasible. Against these preferences is one significant consideration; an important book by a young economist should, in many instances, be reviewed by an established scholar in the subfield simply to help identify a new "creative figure."

I have included in Table 2 brief reference to the costs of putting out the *Journal of Economic Literature* and the forthcoming annual indexes. It is appropriate that the cost problem be discussed most thoroughly with the Executive Committee, since the responsibility for the administration of the Association is vested in the Executive Committee. We have pursued what we believe are cost-effective policies. We anticipate an increase in the printing cost of putting out the *Journal*, particularly because of the expansion of the index of articles section. Also there is the matter of page cost; I should add that the cost of paper rose approximately 40 percent in one month (mid-August to mid-September, 1973).

Table 3 refers to the technical difficulty of the material in the *Journal of Economic*

TABLE 2—EXPENDITURES 1969 THROUGH 1973[a]

	JEL Printing[b]	*JEL/Index* Editorial, Bibliographical, Payments to Contributors, etc.	Total
1969	$104,721	$47,087	$151,808
1970	116,845	63,544	181,090
1971	115,127	74,284	189,410
1972 Budget	117,500	87,500	205,000
1972 Expenditures	118,442	85,083	203,525
1973 Budget	119,950	94,661	214,611
1973 Estimated Expenditures	134,540	90,800	225,340

[a] Figures for 1969–72 are coordinated with the auditor's figures.

[b] Includes mailing, addressing, and gross cost of reprints.

Table 3—Classification by Technical Difficulty

	Surveys	Creative Curmudgeon Articles	Others[a]	Total
Most Difficult	6	7	1	14
Some Difficulty	9	8	6	23
Not Difficult	1	1	19	21
Totals	16	16	26	58

[a] Review articles or books, general essays on all literature.

Literature. From time to time questions have been raised from the floor of the convention relating to the "readability" of the *Journal of Economic Literature.* We strive to produce a journal where the survey articles and the other essays should be relatively easy for economists who have completed one year of graduate training to comprehend. Sometimes our articles are too difficult; in other instances, the subject matter does not require so much expertise. In any event, we believe that the problem requires constant monitoring; G. L. Bach and others have quite appropriately reminded us from time to time that the great majority of the members of the Association must be able to read easily our product; it is a topic that deserves our constant vigilance.

Table 4 is an attempt to show the distribution of material by subject matter. It extends only to the commissioned survey articles and the other essays which we publish. I believe that a quick glance at the distribution of books reviewed reflects the kind of balance which is desirable.

Sometime in the future we will be preparing for an International Economic Association meeting an analysis of the "shift" in interest from one subfield to other subfields as shown by articles published in approximately 250 journals over the past quinquennium. (Somewhere in the neighborhood of 7,000 articles are published per year.)

In the 1974 volume of the *Journal of*

Table 4—Classification by Subject

Classification	Commissioned Survey	Creative Curmudgeon Essays	Other[a]	Total
01 General	1	–	4	5
02 Theory	4	6	1	11
03 Thought (Methodology)	–	–	9	9
04 Economic History	–	–	2	2
05 Comparative Systems	1	1	1	3
11–12 Growth and Development	1	3	–	4
21–22 Econometric, Statistical Theory, Statistics	2	–	1	3
31 Monetary Economics	–	2	1	3
32 Fiscal Economics	1	2	1	4
40–44 International Economics	2	2	5	9
50 Managerial Economics	–	–	–	0
60 Industrial Organization, Industrial Regulation	–	–	–	0
70 Agricultural Economics	1	–	–	1
80 Labor Economics	1	–	–	1
90 Applied Welfare Economics, Regional Economics	2	–	1	3
Totals	16	16	26	58

[a] Review articles on books, general essays on all literature.

Economic Literature we will have four survey articles; we have commissioned and will print four from the following areas—the economics of the black ghetto; recent developments in capital theory, particularly stressing the work done in the neoclassical tradition (in apposition to the earlier Harcourt article, an article which had considerable reaction); the impact of increased women's participation in the economy; the literature on research and development; recent work in government regulation of industry; the economics of racial discrimination in the labor market; and, possibly, some recent work in monetary economics.

We have under consideration interpretive essays on the relationship of economic development and population growth, the corporation and its social role, recent work on expenditure over the life cycle, recent work on fiscal federalism, and several articles dealing with different phases of the relationship of advanced (industrialized) to developing (nonindustrialized) economies. We will also have some articles dealing with some rather interesting work being done in the history of economic thought—particularly dealing with aspects not always available to most of us trained in the English speaking traditions. We will also have some review articles of books.

The Board of Editors has been, as ever, diligent in its labor. Two finish their terms of service with the end of this calendar year; they are Irma Adelman (University of Maryland) and William Baumol (Princeton University and New York University). They have been so helpful on so many occasions that I am nonplussed when it comes to expressing adequately my gratitude. The other members of the Board of Editors similarly assisted me far beyond any right that I had to expect. They are Marcus Alexis (Northwestern University), Abram Bergson (Harvard University), Martin Bronfenbrenner (Duke University), Anne O. Krueger (University of Minnesota), Allan H. Meltzer (Carnegie-Mellon University), Harry T. Oshima (The Rockefeller Foundation and University of the Philippines), and Arnold Zellner (University of Chicago). Finally, I wish to thank a large number of individuals upon whom I have called for specific acts of assistance during this calendar year.

Albert Ando
Joseph Aschheim
John S. Austin
Andrew F. Brimmer
Paul H. Cootner
Joseph Dorfman
Herbert Giersch
John Gould
Einar Hope
H. Houthakker
William Jaffé
Melvyn Krauss
Lawrence Klein
Kelvin Lancaster
Marion J. Levy, Jr.
Thomas Mayer
Oskar Morgenstern
Marc Nerlove
Trout Rader
Joan Robinson
Paul A. Samuelson
Ryuzo Sato
Martin Segal
Henry S. Terrell
Vincent Tarascio
Shigeto Tsuru
Hirofumi Uzawa
Jerome Wells
Charles Z. Wilson

MARK PERLMAN, *Managing Editor*

REPORT OF THE COMMITTEE ON HIRING PRACTICES

The Labor Market for Economists

By KATHLEEN BROOK AND F. RAY MARSHALL*

At the annual meeting of the American Economic Association (*AEA*) in December 1971, it was resolved that

> with a view to a radical improvement of the hiring practices in the economics profession, the President and the Executive Committee appoint a committee to recommend to the next annual meeting a code of procedure including among other reforms an open listing of all employment opportunities.[1]

This action was a consequence of numerous persistent complaints—particularly from job seekers—with regard to imperfections in the labor market for economists. These complaints were stimulated by the apparent growing scarcity of academic jobs for economists during the early 1970's and the fear and frustration resulting from the profession's lack of a reliable job information system and orderly and objective hiring practices. It was not so much that there was a lack of any jobs for new Ph.D.s in economics as it was that job seekers would not be able to find the kinds of jobs they thought they were prepared for. Although jobs apparently did not turn out to be as scarce in 1972 and 1973 as it appeared they would be in 1971, the academic job situation is not as favorable for new Ph.D.s as it was during the 1960's, and the present situation is likely to continue for the foreseeable future.

The roots of the present job problems for new Ph.D.s lie in the 1960's, when academic and government agencies acted on the assumption that the United States faced a critical shortage of college teaching and research personnel and there was excess demand for the Ph.D. in almost all fields. During these years, the job seeker's major problem was to make the best selection among several job offers.[2]

In response to this shortage, graduate programs were expanded rapidly. The yearly production of economics Ph.D.s, for example, tripled between 1960 and 1970, and the production of master's degrees more than doubled.[3] As the supply of Ph.D.s grew, however, it became clear to Allan Cartter and a few others that the academic shortage was being transformed into a surplus.[4] Currently, many entrants into the market are more concerned with locating a job than with selecting among several alternatives. Frustration and disappointment result when new Ph.D.s are compelled to accept positions which do not allow them to utilize fully their training and experience.

As jobs become more scarce, there is increased concern with the "fairness" of the market's operations. Many job seekers have only a vague idea as to the appropriate procedures to follow in finding employment and the problems they should anticipate on entering the market. Each candidate has limited information con-concerning vacancies, and the informality of information channels—particularly in the academic market—leads to the conviction that obtaining a good job depends more on personal contacts than on knowledge or skill. This dissatisfaction with the operation of the labor market has led to demands in many fields, including economics, for an open listing of all employment opportunities. (See Minutes of the Annual Meeting.)

---

* University of Texas at Austin. This is a background paper for the American Economic Association's Committee on Hiring Practices.

[1] Minutes of the Annual Meeting.

[2] See David G. Brown, 1965 and 1967, and Howard D. Marshall. These are studies in the allocation of scarce manpower.

[3] Data was obtained from Lindsey R. Harmon of the National Research Council.

[4] See Cartter 1971a, pp. 132–40, and 1971b, pp. 305–10.

Before proceeding, it should be noted that, although critics became more vocal as demand declined relative to supply, they were not absent during the period of excess demand. In 1964 and 1965, David G. Brown surveyed recent labor market experiences of 103 new faculty members and 50 department chairmen at 18 of the largest Southeastern schools. The median social scientist interviewed in the study had sketchy information of only one other job opening at the time he accepted his current position.[5]

I. Objectives

This paper is designed to help job seekers, employers, and others interested in fair and more efficient hiring practices by providing as much information as possible about the labor market for economists. Since we were interested in getting the information together as quickly as possible, we did not undertake a systematic analysis of the labor market for economists, although we think such an analysis is very much needed. Rather, we pulled together all of the information we could find on this subject and supplemented this information with a questionnaire to 185 economics department chairmen. We conducted numerous personal discussions with economists who had recently gone through the job-seeking process and with other individuals knowledgeable about the labor market for economists. We also surveyed the hiring practices and procedures of the U.S. Employment Service and nine other professional organizations.

The specific objectives of this project were to provide more information for job seekers and a better foundation upon which to base the recommendations of the *AEA*'s Committee on Hiring Practices. However, the authors alone are responsible for any defects in this paper, although the recommendations in the final section are those of the Committee on Hiring Practices.

The remainder of this paper is divided into six sections. The next section discusses the characteristics of economists in the United States—their numbers, educational backgrounds, types of employment, primary activities, salaries, etc.—and is based primarily on studies conducted by the National Science Foundation (*NSF*). Following this, we consider various submarkets—academia, government, and business and industry—as well as the extent of unemployment among economists. The fourth section deals with mobility between economics submarkets. The fifth section deals with equilibrating adjustments in markets characterized by excess supply or excess demand. The sixth section discusses the job market activities of several professional organizations. The last section presents the *AEA* Committee on Hiring Practices' recommendations for improving the operation of the labor market for economists.

II. Characteristics of Economists

Estimates of the number of economists

---

[5] See Brown 1965, p. 261. Only those schools with departments of economics, sociology, and history with at least three members each were included in the sample.

in the United States vary widely, depending on the definition of economist and the method of surveying the population. For example, the 1970 census reported 67,000 economists (see U.S. Dept. of Commerce), while the Bureau of Labor Statistics (*BLS*) reported only 33,000 economists (see U.S. Dept. of Labor). The census provides no detailed individual characteristics, but the *BLS* reported that over half of all economists were employed by business and industry, over one-fourth by colleges and universities, and about one-sixth by government agencies. Since a bachelor's degree is sufficient for many beginning research jobs in government and industry, it would appear that these data are more heavily weighted toward holders of the baccalaureate degree than those reported by the *NSF*.

The most complete information available on the characteristics of economists is that compiled by the *NSF* from the National Register of Scientific and Technical Personnel. Scientists were considered eligible for inclusion in the National Register if they had "full professional standing" as determined by the appropriate scientific professional society—in this case, the *AEA*.[6] Since our interest is mainly in the market for economists with postgraduate training and since over 90 percent of the economists in the National Register have graduate degrees, we have relied primarily on the *NSF* data.[7]

In 1970, 13,386 economists were registered with the *NSF* (4 percent of all scientists registered), and they had a median annual salary of $16,300. Of these, 812 (6 percent) were women; their median annual salary was $13,400.[8] In terms of median annual salary, economists ranked third among all scientific fields; only statisticians and computer scientists were higher. Table 1 provides some idea of the recent growth of the profession.

TABLE 1—ENROLLMENT FOR ADVANCED DEGREES

Year	Economics (excluding agricultural economics)	Agricultural Economics
1960	5,076	1,081
1962	6,429	1,371
1964	8,141	1,504
1966	9,005	1,534
1968	10,585	1,791
1970	11,507	1,781

Source: American Council on Education.

Economics and agricultural economics degrees conferred in the 1969–70 academic year were as follows:[9]

	Economics (excluding agricultural economics)	Agricultural Economics
Bachelor's	17,528	1,067
Master's	1,990	406
Doctorate	794	181

Salary and type of employment vary with the degree held. In 1970, 54 percent of all economists held Ph.D.s, 37 percent held master's degrees, and only 9 percent were in all other categories. The salaries for those with Ph.D., master's, and bachelor's degrees, respectively, were

---

[6] An economist is defined as an individual with a graduate degree in economics, or a bachelor's degree in economics plus two years of professional experience, or the equivalent in professional experience. For the *NSF* survey, the *AEA* developed a comprehensive mailing list from membership lists, lists of nonmembers attending professional meetings, publication subscription lists, and alumni organizations.

[7] The most recent data available are contained in *NSF*, *American Science Manpower, 1970* (1972a). Except where specifically noted, all data on the characteristics of economists and their employment were obtained from this source.

[8] See Carolyn Shaw Bell, pp. 134–42. Bell argues that the National Register underreports the number of women scientists since it identifies a scientist as a working scientist and a working scientist as one with paid employment.

[9] See U.S. Office of Education.

TABLE 2—NUMBER OF ECONOMISTS BY PRIMARY WORK ACTIVITY AND MEDIAN ANNUAL SALARY, 1970

Primary Work Activity	Number of Economists	Median Annual Salary
Research and development	2,746	$16,500
Basic research	782	15,800
Applied research	1,840	16,800
Management or administration	2,937	22,000
Management or administration of research and development	1,618	22,500
Teaching	5,288	16,000
Production and inspection	186	16,100
Consulting	463	20,000
Exploration, forecasting, reporting	532	17,000
Other	358	19,000
Not employed	543	
No report of work activity	333	16,500

$17,400, $14,200, and $16,300. Those who held advanced degrees were concentrated in educational institutions, which had relatively low median salaries as compared with the federal government and business and industry where bachelor's level economists more frequently were employed.

Economists were engaged in a wide variety of activities in 1970 but were most commonly found in teaching (40 percent), management or administration (22 percent), and research and development (21 percent). Table 2 contains more complete information on primary work activities and median salaries.

The *NSF* survey is not limited to those individuals who held advanced degrees in economics. Data are available showing the number of economists whose highest degrees were in scientific fields other than economics, as well as the number whose major subject was economics but who were employed in some other scientific field. Of the 13,386 economists listed, 11,997 had economics as the major subject of the highest degree, but 226 had majored in political science, 81 in agricultural science, 72 in statistics, 60 in mathematics, and the remainder in other subjects. On the other hand, 15,561 of the registered scientists held their highest degree in economics. Most (11,997) were economists, although 1,022 were chemists, 917 were mathematicians, 713 were computer scientists, 212 were statisticians, and 167 were agricultural scientists.

Tables 3 and 4 give the major field for doctorate-level economists and the field of employment for scientists with a doctorate in economics. These data indicate that 93 percent of all registered economists who held the doctorate majored in economics and that 95 percent of all scientists holding a doctorate in economics were employed as economists.

The distribution of economists by field of specialization varied with the degree

TABLE 3—NUMBER OF ECONOMISTS HOLDING DOCTORATES, BY MAJOR SUBJECT OF DOCTORATE, 1970

Major Subject of Doctorate	Number of Economists
Economics	6,709
Political science	81
Agricultural science	32
Statistics	20
Engineering	7
Mathematics	6
Other	232
Not reported	138
TOTAL	7,225

TABLE 4—NUMBER OF SCIENTISTS HOLDING ECONOMICS DOCTORATE, BY FIELD OF CURRENT EMPLOYMENT, 1970

Field of Current Employment	Number of Scientists
Economics	6,709
Mathematics	162
Statistics	37
Agricultural science	29
Psychology	24
Sociology	22
Computer science	20
Other	52
TOTAL	7,055

held. Ph.D.s were concentrated most heavily in general economic theory, history, and systems; business organization; monetary and fiscal theory; and economic growth and development. The predominating fields for master's degrees were business organization, agriculture and natural resources, and general economic theory. Table 5 gives the distribution for all economists.

TABLE 5—NUMBER OF ECONOMISTS BY SUBFIELD, 1970

Subfield	Number of Economists
General economics	1,626
Economic growth and development	1,131
Economic statistics	576
Monetary and fiscal theory	1,260
International economics	604
Business organization	2,153
Industrial organization	805
Agriculture and natural resources	1,501
Manpower, labor, and population	828
Welfare, consumer economics, urban and regional economics	1,703
Economics, other	1,954

III. Academic Submarket of the Market for Economists

In the past, about two-thirds of all economics Ph.D.s have been employed in education, either as teachers, researchers, or administrators.[10] One observer has commented that the traditional idea of professional bliss in economics is "employment at a high-status university and a research career only slightly sprinkled with teaching responsibilities." More specifically, "a 'good' economist is a research academician at a big-name university who stresses publication and identifies with other research economists" (Robert P. Strauss, p. 327). However, the prospect for the future academic market is one where many new doctorates who aspire to academic careers at large universities will find themselves teaching at smaller schools or turning to the nonacademic world for employment. Those who teach will find themselves spending more time with undergraduate students and less time doing research than they might prefer.

TABLE 6—NUMBER OF ECONOMISTS EMPLOYED BY EDUCATIONAL INSTITUTIONS, BY HIGHEST DEGREE, 1970

Highest Degree	Number of Economists
Doctorate	5,449
Master's	2,127
Bachelor's	245
Less than bachelor's	1
No report	23
TOTAL	7,845

Three major factors have contributed to this relatively greater supply in the academic market. First, during the next ten to fifteen years, there will be a decline in the rate of increase of the college-age population. Cartter, for example, has predicted declines in student enrollments and in the sizes of university faculties by 1980 (Cartter 1971a). Thus, the number of new faculty needed to maintain the present faculty-student ratio will decline and possibly become negative. Second, there is likely to be a shift in the enrollment mix, with lower-division education—particularly in the junior colleges—growing more rapidly and graduate education more slowly than in the 1960's. This will reduce the demand for new doctoral faculty needed to maintain the quality of the educational system, especially to the extent that junior colleges view the research orientation of the Ph.D. as being incompatible with good lower-division teaching. Third, as a result of the expansion of graduate programs in the 1960's, current faculty members are relatively young. This means that the rate of retirement of faculty members, and hence the demand for replacements, can be expected to decline.

---

[10] See Francis M. Boddy, p. 316. In a survey of fifty-four of the largest economics departments in the country, Boddy found that approximately 80 percent of the 1,100 Ph.D.s who received their degrees between 1966 and 1969 worked in educational institutions.

According to Cartter, the academic market is capable of handling deficits or surpluses of as much as 15 percent of available doctorates. With a small imbalance, most available doctorates are hired by colleges and universities, and the percentage of faculty holding the doctorate adjusts accordingly (Cartter 1971b). But during the 1970's, the anticipated surplus probably will be too large to be absorbed in academia unless faculty-student and Ph.D.-non-Ph.D. ratios are changed significantly.

The operation of the academic labor market has been analyzed fairly extensively by Brown and Marshall. Although both men conducted their research during the early 1960's, when supply of college and university teachers was smaller relative to demand, a number of generalizations are pertinent:

(1) Academic placements display a "downstream" pattern, i.e., most job candidates should expect to find placement in departments with the same or lower quality ratings than their own graduate department. The main basis for judging the potential of a new Ph.D. tends to be the prestige of the department that granted his degree. Top-ranked departments supply faculty members to each other and to lower-ranked departments, but obtain relatively few faculty members from the lower ranks.[11]

(2) Many schools are reluctant to hire their own Ph.D.s immediately after they obtain their degrees, but not after they have been broadened by teaching experience elsewhere. However, there is a tendency for some schools to favor their own Ph.D.s because they can generally be recruited with minimal expense and effort and because they are better known than outsiders.

(3) Because of budget considerations and faculty morale, most hiring occurs at the level of instructor or assistant professor.

(4) Schools are more willing to compromise on the desired characteristics of candidates than on salaries.

(5) A number of professors enter at least the fringes of the labor market with little intention of actually changing jobs but desiring to establish their current market values.

These studies also provide some information on the relative importance of formal and informal methods in locating jobs. Brown, studying eighteen major institutions in the Southeast, found that informal methods were most common. Thirty-four percent of the social scientists interviewed learned about their current jobs from friends by letter or telephone, 12 percent from friends at the convention, 23 percent through the graduate school department, and 11 percent through a major graduate school professor (Brown 1965, p. 102). Marshall's study of economists yielded similar results, with greatest reliance being placed on friends, informal contacts at annual meetings, and the chairman of the graduate department (Marshall, p. 85).

On the other hand, relatively few social scientists located employment through more formal channels—7 percent through mass letters of application, 4 percent through university and college placement services, 2 percent through ads in professional journals, and 1 percent through commercial teacher placement agencies. Brown notes that even those who participated in formal clearinghouses had little faith in them. Listing one's name was regarded as a "job-finding insurance"—a way of ensuring that some job would be found, rather than a way of finding the best job available. Marshall and Brown suggest that small schools probably make greater use of formal agents than do large schools, although Marshall observes that a number of the larger colleges in the South and Midwest made regular use of private placement bureaus (Marshall, p. 82). Brown (1967) notes that

---

[11] See John R. Niland, pp. 141–56. In this article, Niland uses the proportion of new Ph.D.s taking employment at a university of equal or better quality as a measure of the tightness of academic labor markets.

> Out of all formal methods, the convention placement service is *least* often considered "unprofessional" or "worthless" and is the *most* often not used because of a "lack of contacts" (didn't attend the convention). . . . [O]f all formal methods, it turns up the greatest percentage of jobs in the top salary range, the greatest percentage of jobs in top prestige schools, and the second greatest percentage with the lowest teaching load. [p. 132]

Economists have three main formal placement channels:

(1) In each issue of the *American Economic Review* (except *Papers and Proceedings*), the *AEA* publishes, without charge, descriptions of vacancies and applications submitted through the Secretary of the Association. Announcements are repeated for a maximum of four issues in two years and can be made anonymously. The Association plays no role in negotiations between employers and job candidates; it merely acts as a central point for forwarding inquiries. In his study, Marshall found little use of these advertisements by employers. The results of our survey (discussed in detail below) confirm that a sizable minority of schools use ads in professional journals but that most of them do not.

(2) The National Registry for Economists, established in 1966, provides a nationwide year-round clearinghouse for economists. This is a free service operated by the Chicago Professional Placement Office of the Illinois State Employment Service. Applicants are matched with vacancies by computer on the basis of qualifications and interests. The application forms of the selected candidates are then forwarded to prospective employers. Resumés are also accepted by the National Registry but are forwarded to employers only upon request. Most of the registrants are Ph.D. candidates or holders, and most vacancies are in academic institutions and therefore require a Ph.D.

(3) The U.S. Employment Service (*USES*) operates job placement facilities at the *AEA* annual meetings. At the 1971 meeting, there were 719 job openings and 1,791 job applications (see Report of the Secretary, p. 480). At the 1972 meetings, there were 1,586 applicants and 650 prospective employers.[12] All vacancies and applications listed with the National Registry are available at the placement center. Applicants and employers may review the entire file and arrange personal interviews.

Most academic jobs are obtained through informal channels, particularly through friends and graduate school professors. One such channel in economics is the Chairmen's Group (*CG*), an informal group composed of forty-three university departments in the summer of 1973, later expanded to over sixty, which award about two-thirds of all Ph.D. degrees in economics.[13] These departments exchange information on candidates, job openings, and salary schedules. Lists of graduate students who will be seeking jobs are consolidated and distributed to each school.

The results of the survey conducted in connection with this paper, presented in Table 7, indicate that while only about

---

[12] Data was obtained from Robert Brown, U.S. Employment Service.

[13] Appendix A contains a list of the *CG* schools.

TABLE 7—RECRUITING TECHNIQUES OF DEPARTMENT CHAIRMEN USING THEM REGULARLY OR OCCASIONALLY
(By Percent)

Recruiting Technique	*CG* Ph.D.	Other Ph.D.	Master's	Bachelor's
University placement services	32	57	71	62
Private placement services	6	12	14	10
Ads in professional journals	18	35	28	34
Professional meetings, informal contacts	94	82	100	79
Professional meetings, *USES*	18	30	50	27
Personal contacts with friends and other department chairmen	97	90	93	90
Blind letters of applications	41	67	86	76
Letters of recommendation	88	82	100	72

one-fifth of the *CG* schools use the *USES*, this was a more important hiring channel for the schools in the other categories.

There have been a number of criticisms of the *USES* operation at the *AEA* conventions. A common complaint from job seekers is that they must rush around frantically during the first few hours of interviewing, trying to arrange appointments. It has therefore been suggested that interview schedules be set up prior to the opening of the annual meetings in order to make this process more orderly. In addition, the physical facilities often have been inadequate—in some cases with several interviews being conducted simultaneously in the same room. The *AEA* should therefore give adequate physical facilities for interviewing a high priority in making arrangements for annual meetings.

Other formal channels are open to all individuals seeking academic positions—not just economists. For example, for a nominal charge, the *Bulletin* of the American Association of University Professors lists vacancies and teachers available, and the American Association of University Women maintains a roster of women holding earned doctorates for use by educational institutions in locating staff.

Examples of placement services operated by church groups are the Cooperative College Registry and the placement office of the Board of Education of the United Methodist Church (*UMC*). The latter office serves eight universities and ninety-seven four-year colleges. The *UMC* agency refers applications from its files to employing institutions on request. Unfortunately, there are no data on the number of such placements, since the office takes no part in negotiations between candidates and schools. The director stated that member schools probably rely more on professional associations and less formal contacts than on the *UMC* service. The Cooperative College Registry represents the recruitment efforts of more than 200 colleges sponsored by ten Protestant denominations. Representatives visit campuses of over 120 graduate schools and 25 national conventions to register prospective candidates.

Most colleges have their own placement offices which maintain permanent records of graduate school performance, including letters of recommendation. Brown (1967, p. 128) found that these offices were more likely to be consulted by employers in disciplines with short supplies than those where applicants are more plentiful and that only 3 percent of the placements assisted by these offices were in top quintile schools.

Finally, several commercial agencies make academic placements, including the College and Specialist Bureau and the American College Bureau, which provide credential referral and locator services and make over 100 college-level placements each year. Most of the jobs listed are average or below average and at the beginning academic ranks. Most of the candidates registered are experienced. The fee charged if placed is usually about 5 percent of the first year's salary (Brown 1967, p. 133).

One of the most popular job-seeking methods is the blind letter. Brown (1967, p. 124) states, "One of the most surprising results of the study is the frequent and successful use of the direct letter of application, the letter sent cold to a prospective employer." According to his study, 19 percent of all teachers found the jobs they actually accepted by this means. Attitudes toward blind letters vary. Some recruiters prefer more subtle approaches or prefer to take the initiative themselves. Marshall (p. 133) suggests that it would be useful for departments which disregard letters of application to publicize this fact.

While blind letters may be an effective method of obtaining employment, Marshall thought applicants needed to improve both their selection of colleges and the quality of their letters. This is particularly necessary in a buyer's market. Panic may drive the applicant to send out a mass of letters. An observer of this phenomenon in the case of mathematicians concluded that "it is doubtful that a person who sends out more than 30 letters of inquiry takes the time to determine whether or not he would be willing to accept offers at each of these places, even if they were to be made" (see B. E. Rhoades, p. 389). Applicants should also be aware that chairmen may use rather crude methods of weeding through letters—such as eliminating all mimeographed form letters or all letters not addressed to a faculty member by name.

Additional insights into the importance of various hiring channels is afforded by the questionnaire we mailed to 185 economics departments in the fall of 1973. The questionnaire was pretested on 16 chairmen in August and, after minor revisions, was mailed to the remaining 169 chairmen. The group sampled included all 43 members of the *CG*, 51 other schools granting the Ph.D. in economics, and 10 percent samples of schools offering the master's and bachelor's as their highest degree. Replies received as of November 24, 1973, were as follows:

	Number of Replies	*Questionnaires Mailed*
CG, Ph.D.	34	43
Ph.D.	40	51
Master's but no Ph.D.	14	28
Bachelor's but no graduate work	30	62

The results of this questionnaire with respect to recruiting techniques are presented in Table 7. All schools make primary use of personal and informal contacts, but non-*CG* schools rely more heavily on other recruiting techniques. These schools also give much greater attention to blind letters of application from job candidates. All categories of institutions considered letters of recommendation to be useful, although the bachelor's schools give relatively less weight to letters of recommendation than those in the other categories.

Since most department chairmen recruit at the *AEA* annual meetings, the survey sought to determine whether the meetings are appropriately timed for recruiting purposes. As shown in Table 8, most chairmen are satisfied with the timing of the meetings. Most of those who objected did so because their budgets were not complete by December.

TABLE 8—ATTITUDES OF ECONOMICS DEPARTMENT CHAIRMEN TOWARD THE TIMING OF THE *AEA* ANNUAL MEETINGS

	CG Ph.D.	Other Ph.D.	Master's	Bachelor's
The annual meetings of the *AEA* are appropriately timed for recruiting purposes				
Percent Agreeing	76	80	64	55
Percent Disagreeing	24	15	29	35
Percent Not responding	0	5	7	10

TABLE 9—NUMBER OF GOVERNMENT ECONOMISTS BY HIGHEST DEGREE, 1970

Highest Degree	Federal Government	Other Government
Doctorate	509	277
Master's	724	307
Bachelor's	270	96
Less than bachelor's	2	0
No report	8	4
TOTAL	1,513	684

IV. Nonacademic Submarket of the Market for Economists

Approximately one-third of all Ph.D.s in economics are engaged in nonacademic occupations. Boddy's survey (p. 316) shows that 8.4 percent of those receiving their doctorates between 1966 and 1969 found employment with the *U.S.* government and the Federal Reserve System, 3.5 percent with *U.S.* research and consulting firms, 2.2 percent each with *U.S.* banking, *U.S.* industry, and international organizations, and 4.4 percent with foreign banking, industry, and government.

Next to educational institutions, governments employ the largest number of registered economists—a total of 2,197, of whom about 69 percent are engaged by the federal government (see Table 9). Government economists participate in a wide range of activities, but are concentrated in applied research and research and development management. Those in management received the highest median salaries (see Table 10).

Information on federal jobs is most commonly obtained through direct inquiry or through recruitment by the employing agencies. General information may be obtained from the U.S. Civil Service Commission, which may then direct the applicant to take a civil service examination (required for some positions) or to contact specific agencies which employ economists. Federal job information centers in most major cities provide job announcements, application forms, and pamphlets on request. Similar approaches are used by state and local governments.

TABLE 10—MEDIAN ANNUAL SALARIES OF GOVERNMENT ECONOMISTS

Primary Work Activity	Federal Government	Other Government
Research and development	\$17,000	\$16,400
Basic research	15,800	16,000
Applied research	17,300	16,900
Management or administration	23,000	22,000
Management or administration of research and development	23,000	22,000
Consulting	22,000	—
Exploration, forecasting, reporting	16,900	15,200
Other	20,000	16,400
No report	21,000	—

Employment information for international agencies can be secured from such organizations as the United Nations Recruitment Section in New York, the Food and Agriculture Organization in Rome, the International Bank for Reconstruction and Development in Washington, D.C., and the Organization of American States in Washington, D.C.

Business and industry is the third largest employment category for registered economists. Of 1,822 business and industrial economists, 1,014 (6 percent) hold master's degrees, 466 (26 percent) hold doctorates, and 323 (18 percent) hold bachelor's degrees. The magnitude of research and development activities is a key determinant of the demand for Ph.D.s in industry (see Michael F. Crowley, pp. 65–6).

The demand for business economists has increased during the last two decades, and it is expected to grow at a rate of 5 to 6 percent for the next ten years (see Michael William Stebbins, p. 64). According to *Business Week*, before the late 1950's, "most corporate economists toiled anonymously in market research departments with little noticeable impact on management" (p. 74), but today—as the planning function becomes increasingly vital—economists in business are gaining more recognition.

Business economists work at a wide variety of activities, especially economic forecasting, analyzing the relationship between economic indicators and a firm's performance, and explaining the economic situation to the firm's customers and stockholders. Large firms are conducting simulations of industry and firm performance under changing conditions. Economists with strong computer backgrounds appear to be in particular demand as well as Ph.D.s working on tax, pollution, and government policy problems of business and industry (Stebbins, p. 81).

Table 11—Median Annual Salaries of Business and Industrial Economists by Primary Work Activity, 1970

Primary Work Activity	Median Annual Salary
Research and development	$18,000
Basic research	15,000
Applied research	18,000
Management or administration	23,400
Management or administration of research and development	23,500
Production and inspection	16,200
Consulting	21,000
Exploration, forecasting, reporting	17,800
Other	18,900
No report	16,000

Information on employment in private industry and business may be obtained by contacting the personnel manager of the firm under consideration. In addition, some firms interested in employing economists with advanced degrees send personnel officers on recruiting trips to college campuses. The three placement services discussed in the preceding section (advertisements in the *American Economic Review*, National Registry, and the *USES* facilities at the annual meetings) also are available to economists seeking nonacademic employment.

The remainder of the economists listed in the National Register are employed by nonprofit organizations (4 percent) or by the military (1 percent), are self-employed (1.6 percent), are engaged by some type of employer not specifically listed (.35 percent), or are not employed (4 percent). The median salary of those working for nonprofit organizations is $20,000, as is that of the self-employed. (Salary information is not available for economists working for the military.)

V. Unemployment

A total of 543 economists were not employed at the time of the *NSF* survey—

TABLE 12—NUMBER OF ECONOMISTS NOT EMPLOYED, BY HIGHEST DEGREE, 1970

Degree Level	Total	Seeking Employment	Not Seeking Employment	Retired
Doctorate	127	30	22	75
Master's	371	126	203	42
Bachelor's	41	18	17	6
Less than bachelor's	0	0	0	0
No report	4	0	1	3

174 were seeking employment, 243 were not seeking employment, 126 were retired. Table 12 gives the distribution of these economists by degree level.

In the spring of 1971, when the *NSF* obtained more detailed unemployment information about scientists, the unemployment rate for Ph.D. economists was .5 percent, ranging from a high of 2.5 percent for those over 70 years of age to a low of .2 percent for those aged 40–49 and 54–59 years. The rate for master's-level economists was 3.2 percent, ranging from 12.5 percent for those over 70 years of age to .9 percent for those ages 55–59 and less than .1 percent for those 65–69 years of age (*NSF* 1972b, p. 150). Unemployment rates were also computed by field of specialization—lowest (.7 percent) for those specializing in agriculture and natural resources and highest (2.2 percent) for those in industrial organization (*NSF* 1972b, p. 131).

However, it is generally recognized that unemployment rates are very imperfect measures of the looseness or tightness of the market for highly skilled personnel. "High level manpower tends to be mobile and thus has the ability, mainly by working in lesser capacities, to disguise or at least delay detection of a loosening market" (Niland, p. 141). And according to Cartter, "The price of a surplus of Ph.D.s (surplus in terms of traditional jobs) is not unemployment, but rising underemployment" (1971a, p. 138).

VI. Mobility Between Submarkets

Strauss observed that

> resistance on the part of the new Ph.D.s to nonacademic first jobs seems to come from two sources: fear that the trip out of academe is a one-way journey and a suspicion that gaining an audience for one's research in general economics journals (such as the *AER*) requires an academic affiliation. [p. 328]

He goes on to suggest that this tends to keep capable young people in the academic world and to create a technological and institutional gulf "between those who make policy and those with theoretical and methodological expertise."

Marshall's study of the mobility of college faculties tends to contradict the first assertion in that it indicates considerable movement by economists back and forth between the academic world and business and government. Of the 420 academic economists surveyed, slightly more than 25 percent had worked at some time for state or federal government on a full-time basis, and over 40 percent had experience in the business field, also on a full-time basis (p. 52).

Likewise, a survey by Gerald G. Somers of 280 economists showed that 90 percent had experience in colleges and universities and 58 percent had held nonacademic employment. Of those in academia at the time of the survey, 59 percent had always been in academic work, while 19 percent of those outside academia had always been

TABLE 13—UNEMPLOYMENT RATES OF SCIENTISTS, 1970 AND 1971

(In Percent)

Scientists	Spring, 1970	Spring, 1971
Economists	1.4	1.6
Sociologists	8.6	3.8
Political scientists	3.0	3.4
Anthropologists	0.6	1.3
Mathematicians	1.5	2.6
Chemists	1.5	3.0
Physicists	2.3	3.9
Psychologists	1.0	1.6

Source: National Science Foundation 1972b, p. 2.

in nonacademic work (pp. 509–10). However, both of the surveys referred to were conducted when economists with postgraduate degrees were in short supply. Consequently, mobility may have been much greater than in the market of the 1970's.

In an attempt to gain more current knowledge about mobility, we asked the department chairmen to state the sources of their new faculty and the destinations of those departing their faculties. Their responses appear in Tables 14 and 15.

The majority of all new faculty members for 1973–74 were obtained from other educational institutions—i.e., they were already faculty members or they were attending graduate school. The proportions ranged from a low of 64 percent for schools offering a master's degree but not a Ph.D. to a high of 100 percent for the bachelor's schools. The movement from nonacademic employment in the United States and from employment outside the United States was greatest for schools in the master's category and most restricted for those at the bachelor's level.

On the other hand, much smaller proportions of the departing faculty remained in the academic world. About half of those leaving *CG* departments remained faculty members, while only 6 percent of those leaving bachelor's departments remained in educational institutions as faculty members. However, 18 percent of those leaving master's departments and 12 percent of those leaving bachelor's departments returned to graduate school. Only four of the economists who left all of the schools surveyed were reported to be unemployed.

There is some evidence (cited by Strauss) that nonacademic economists encounter publication difficulties. George J.

TABLE 14—SOURCES OF NEW FACULTY HIRED FOR 1973–74 AS REPORTED BY ECONOMICS DEPARTMENT CHAIRMEN

(In Percent)

	CG Ph.D.	Other Ph.D.	Master's	Bachelor's
Educational institutions in United States, including students	82.3	87.9	64.3	100
Educational institutions in United States, students only	45.8	47.3	21.4	50
Business and industry in United States	1.0	2.2	7.1	0
Government in United States	3.1	3.3	0.0	0
Outside the United States	1.0	2.2	21.4	0
Unemployed	0.0	1.1	0.0	0
Other (total)	12.5	3.3	7.1	0
Research organization	3.1	1.1	0.0	0
Another department at same institution	2.1	0.0	7.1	0
Postdoctoral fellow	0.0	1.1	0.0	0
No information available	7.3	1.1	0.0	0

TABLE 15—DESTINATIONS OF FACULTY DEPARTING AFTER 1972–73 AS REPORTED BY ECONOMICS DEPARTMENT CHAIRMEN

(In Percent)

	CG Ph.D.	Other Ph.D.	Master's	Bachelor's
Educational institutions in United States, as faculty	50.5	37.3	23.5	5.9
Business and industry in United States	4.0	14.7	17.7	17.7
Government in United States	12.1	9.3	0.0	11.8
Left the United States	12.1	10.7	5.9	5.9
Death or retirement	13.1	14.7	17.7	5.9
Still seeking employment	0.0	4.0	5.9	0.0
Other, total	8.1	9.3	29.4	52.9
Research organization	2.0	0.0	0.0	0.0
Another department at same institution	2.0	2.7	0.0	5.9
Return to graduate school	3.0	1.3	17.7	11.8
Housewife	1.0	0.0	0.0	0.0
International Monetary Fund	0.0	0.0	5.9	0.0
No information	0.0	5.3	5.9	35.3

Stigler has found that "apparently governmental service is unfavorable to participation in the main issues and controversies of professional economics" (p. 45). Government economists publish primarily in government journals. There is also an observed tendency for authors who have received degrees from a given institution to be disproportionately represented in the journals published by that institution (Michael C. Lovell, p. 44). Preserving anonymity in the review process might alter this situation.

The mobility of economists between academia, business, and government, as well as within each sector, deserves much more study than it has received, because it is of great concern to all of those interested in the labor market for economists, but especially to those who wish to know whether their first jobs will determine the courses of their careers.

VII. Equilibrating Adjustments

We now turn to the nature of the equilibrating adjustments in the labor market for economists, especially in view of the relative excess supplies likely to prevail for the foreseeable future. Boddy has already noted some price adjustment. His figures indicate a smaller rise in the hiring salaries for new Ph.D.s in 1970–71 than in previous years. He also notes that statements by departments concerning what they expect to pay in the future show little evidence of any continued rise (p. 318).

The data in Table 13 show that there has also been a slight rise in unemployment rates among economists and other scientists.

One of the most obvious solutions to the problem would seem to be a reduction in the supply of economists, accomplished by reducing the number of advanced degrees granted in the field. Toward this end, Harvard and Yale in 1971 announced decisions to reduce incoming graduate classes by about 25 percent, and New York State announced a one-year moratorium on the approval of new Ph.D. programs (Ted R. Vaughn and Gideon Sjoberg, p. 147). Our survey indicates that a sizable proportion of the economics departments offering the Ph.D. have reduced their graduate enrollment (see Table 16). Among the *CG* departments, this resulted from a decision to

TABLE 16—PERCENTAGE OF DEPARTMENTS WITH INCREASING, DECREASING, OR UNCHANGED GRADUATE ENROLLMENT (1969–70 TO 1972–73)

(In Percent)

	CG Ph.D.	Other Ph.D.	Master's
Increase	17.7	17.5	28.6
Decrease	38.2	45.0	14.3
Unchanged	41.2	35.0	50.0
No response	2.9	2.5	7.1

reduce the number of graduate students by raising admissions standards. In some departments, the reductions resulted from budget cuts which brought about decreases in financial support for graduate students. In other cases, department chairmen suggest that prospective graduate students were discouraged by the present job market situation. However, Cartter anticipates that such reductions will be more than compensated for by expansions in graduate enrollment at universities which are just in the process of developing Ph.D. programs. In addition, there is a lag of at least four years, and probably more, between the time a decision is made and the time its effects become apparent in the marketplace.

A more probable adjustment is the upgrading of economists in government, industry, and less prestigious educational institutions. That is, there will be an increase in the proportion of economists in these institutions who hold doctorates as well as an increase in the proportion who received their doctorates from top-ranking departments. Niland argues that this "trickle-down" process would check the trend toward increased quality variance which was a consequence of rapid expansion of Ph.D. programs during the 1950's and 1960's and would "reestablish greater homogeneity within and between conventional specialty designations" (p. 152). Nevertheless, such benefits would be accompanied by considerable personal disappointment to those who had expectations of finding placement at major universities and achieving "professional bliss" unless, of course, narrowing the quality gap between institutions reduces the distinction between "prestige" jobs and all others.

The chairmen of departments offering the Ph.D. indicated to us that there has already been some decline in the quality of employment obtained by their new Ph.D.s and that they are encouraging their graduates to look into types of employment which would have received little consideration in previous years.

The survey also shows that master's and bachelor's level departments are currently substituting Ph.D.s for non-Ph.D.s. Only 35 percent of the faculty members who left master's departments in 1972–73 held the doctorate while 86 percent of those who were hired held that degree. The corresponding figures for the bachelor's departments were 47 percent and 64 percent respectively.

In some portions of the academic labor market, especially in physics and mathematics, movement toward equilibrium is occurring at the expense of the foreign component in *U.S.* supply. For example, new Ph.D.s are replacing foreign doctorates as well as some predoctoral students in positions as research and teaching assistants (Cartter 1971a, p. 138). As noted in Tables 14 and 15, a larger proportion of faculty members leaving *U.S.* academic institutions are finding employment outside the United States than the percentage of new hires who come from abroad.

Turning now to the demand side of the market, it has been suggested that professional associations could encourage the employment of Ph.D.s in relatively new areas, such as two-year colleges, where their research orientation has been suspect. Such efforts might indicate a need for changes in the nature of the Ph.D. product. Some observers of the market for Ph.D.s assert that graduate training has become increasingly narrow, so that students are prepared only for entrance into academic life at the university level. One mathematician states that he finds little evidence that junior colleges or industry are eager for people like the majority of Ph.D.s being produced (Gail S. Young, pp. 718–22). Another, Reese T. Prosser, argues that

> the graduate student with a broad background in modern mathematics, a strong interest and sound training in undergraduate teaching, and a nodding acquaintance with statistics, computers, or graphs has a formidable advantage over the graduate with a single research specialty, no matter how good he is, in today's market. [p. 615]

To be more specific, it would be beneficial to students if graduate programs would place greater emphasis on the applications of knowledge.

A unique problem for academia arising from present demand and supply conditions should be noted (Joseph P. LaSalle):

> The transition to a "steady state," since our faculties are now relatively young, will create a real crunch for younger untenured members of the faculty and will make it difficult for us to maintain that balance between younger and older professors so necessary for a vigorous department and for lively teaching. [p. 69]

Early tenure decisions may severely limit the ability of new Ph.D.s to obtain academic positions. A possible solution is to devise some sort of intermediate arrangement at the end of six years' experience, for example, delaying a permanent commitment until the completion of twelve years of service (Cartter 1971a, p. 139). A reduction in the age of retirement would also increase openings to some extent, although not enough to absorb the entire surplus.[14]

VIII. Job Market Activities of Professional Organizations

During the 1970's, almost all academic disciplines face the problem of having relatively fewer jobs for graduates than was the case during the 1960's. A number of professional organizations therefore have adopted measures to improve their labor markets. In order to get some suggestions for possible ways to improve the operation of the market for economists, we

TABLE 17—PROPORTIONS OF ECONOMICS FACULTY WITH AND WITHOUT THE PH.D. AS REPORTED BY DEPARTMENT CHAIRMEN

(In Percent)

	CG Ph.D.	Other Ph.D.	Master's	Bachelor's
1973–74 Faculty				
With Ph.D.	93.9	89.6	71.0	66.9
Without Ph.D.	6.1	10.4	28.9	33.1
Faculty departing after 1972–73				
With Ph.D.	84.8	73.3	35.3	47.1
Without Ph.D.	15.2	26.7	64.7	52.9
New faculty hired for 1973–74				
With Ph.D.	77.1	65.9	85.7	64.3
Without Ph.D.	22.9	34.1	14.3	35.7

---

[14] An investigation of financial and other aspects of early retirement is necessary in evaluating the feasibility of this approach.

examined the job market activities of ten professional organizations.

(1) The *AEA* in the past has played a rather passive role in the job market, but, as noted above, it does make available three formal placement services to its 18,000 members. These are the advertisements in the *American Economic Review*, the National Registry for Economists, and the U.S. Employment Service facilities at the annual meetings. These have already been discussed.

(2) One proposed solution to problems in the job market is the open listing of all employment opportunities. The American Mathematical Society (*AMS*), with its 13,000 members, is one of the groups that has undertaken such a project. In October 1973 it published the first issue of *Employment Information for Mathematicians*, containing four sections— (a) a list of all available positions submitted by department chairmen, (b) a list of all institutions which have no available positions, (c) a list of all institutions which failed to respond to inquiries, and (d) information on government, industrial, and foreign positions as received by the *AMS*. Information contained in the publication is derived from a bimonthly postcard survey of department chairmen. The lists will be published in February, April, June, July, October, and December at subscription rates of $20 per year. In addition to the job lists, the society conducts an annual salary survey and a survey of faculty mobility.[15]

(3) Physicists have faced a particularly acute job crisis. The American Institute of Physics, a federation of seven national societies with a total of 53,000 members, has established a Manpower Division which surveys the employment of new B.S. and Ph.D. recipients in physics. It also has initiated surveys to gather data on the physics profession formerly collected by the *NSF* National Register of Scientific and Technical Personnel.

(4) A group of young physicists who are concerned with improving the economic and professional standing of physicists has established the American Physicists Association, which has just completed a comprehensive survey of the employment situation.

(5) The American Political Science Association (*APSA*), with 15,000 members, mails a monthly personnel services newsletter listing job openings to subscribers for $6 per year. The association considers the open listing policy to have been successful and has experienced a sharp increase in the total number of listings—from 382 in 1971–72 to 523 in 1972–73 (see Rona B. Hitlin). The *APSA* also operates a credential referral service. Files which include resumé, dissertation abstract, letters of reference, and lists of publications are referred on request by a member or institution. The charge is $8 per year. Last, the *APSA* has established an Organization of Placement Directors. This group, during its brief existence, has conducted a survey of placement success and held an open meeting at the last *APSA* convention.

(6) In January 1973 the National Association of Business Economists (*NABE*), with a membership of about 2,000, appointed a special committee charged with improving the job market for business economists. Subsequently, a private placement agency, David Cordell, Inc., was chosen to operate the *NABE* placement service. *NABE* also conducts a salary survey of business economists every four years.

(7) The 34,000 members of the American Psychological Association (*APA*) are served by a monthly employment bulletin which lists positions and availability notices. The latter appear anonymously. Subscriptions are $7 per year, and the circulation is about 5,000. Position openings

---

[15] See *American Mathematical Monthly*, p. 115.

also appear in the association's newsletter. Annual surveys are conducted regarding the employment status of new doctorates and the use of placement services at regional conventions. The most noteworthy effort of the *APA* is the compilation of a manpower data system, designed in part to fill the void left by the demise of the *NSF* National Register of Scientific and Technical Personnel. The data are to be used for biographical publications and statistical analyses of the characteristics of psychologists and to identify pools of individuals with particular expertise. It has been suggested that the system be expanded to include up-to-date information on hiring and job placement, but as yet there is no indication that the *APA* is moving in that direction.

(8) The 30,000-member Modern Language Association (*MLA*) has achieved some success in publishing job information lists. In a recent survey, more subscribers said they found new positions through the lists than by any other means (47.1 percent of those finding jobs in English and 45.9 percent in foreign languages).[16]

(9) In September 1972 the American Historical Association (*AHA*), with 18,000 members, announced the adoption of the *MLA* information system. Simultaneously, it abandoned the Professional Register operation at the annual convention. (The service had offered few concrete results at great expenditure in money and organizational effort.) Under the *AHA* system, the staff takes the initiative in obtaining information about current vacancies and the general hiring situation in departments. The cost is approximately $28,000 per year, but is offset by charging sufficient fees to cover costs.[17]

(10) The American Sociological Association, with about 13,500 members, publishes a newsletter containing vacancy and availability notices and an annual job market report describing the annual production of Ph.D.s, enrollment adjustments, and the expected number of vacancies.

IX. Job Listings

Two types of procedures have been followed in the publication of job lists. Some groups have assumed the initiative in obtaining information about current vacancies by conducting periodic surveys of department chairmen. Others merely accept and publish advertisements submitted at the discretion of employers. In both cases, the professional association's responsibility ends with the publication of the lists, since it takes no part in negotiations between candidate and employer.

Several problems are presented by the publication of job lists: (1) there has been a tendency for the more prestigious departments not to list their openings; (2) there is a lag between the submission of the listing and the publication and mailing of the list; and (3) departments receive an avalanche of letters in response to a listing. Thus some guidelines must be established by the association if the listings are to be successful.

Since the *APSA* is approximately the same size as the *AEA* and has published job listings for a number of years, its experience might be useful to the *AEA*.

In 1970 the Council of the *APSA* adopted the following resolution:[18]

> It is a professional obligation to all political science departments to list in the *APSA* Personnel Service Newsletter all positions for which they are recruiting except those vacancies at the Associate and Full Professor levels which departments expect to fill from among people known to them. [p. 73]

---

[16] See Modern Language Association *Newsletter*, p. 3.

[17] See American Historical Association *Newsletter*, p. 4.

[18] See American Political Science Association *Council Minutes*, p. 73.

Since the *APSA* took this position, there has been a substantial increase in listings and in the number of prestigious institutions which have listed. During the 1972–73 academic year, fifteen of the top twenty departments listed their vacancies. However, it is probably unrealistic to assume that this has brought about a major change in hiring practices among these departments. Furthermore, the *APSA* indicates that the increase may be attributable to *HEW* affirmative action guidelines, since public listing of vacancies provides a vehicle by which departments may comply with these directives.

The newsletter, which is published monthly, is typed at the national office in Washington, D.C., and then photocopied. Listings are added until the very last moment, but there is a four- or five-day lag between typing and mailing.

Guidelines for the use of the service include the following: (1) departments should provide specific job descriptions to avoid an avalanche of letters; (2) they should not fill a vacancy for at least thirty days after the publication of the notice; (3) they should acknowledge each application, if only by means of a form letter; (4) candidates should apply only for positions for which they qualify; and (5) they should include a postcard with their applications so that departments may respond with minimum cost and effort.

Over 1,200 individuals and 400 departments pay $6 per year to receive the *APSA Newsletter*. Approximately half of the individual subscribers are annual rather than student members of the association. For the last two years, a questionnaire has been included in the summer issues to determine the effectiveness of the newsletter in job search. The newsletter was most frequently mentioned as the respondents' primary source of leads.

Our survey provided information on the attitudes of economics department chairmen toward the usefulness of open listing and a computerized listing of job candidates (see Table 18). Probably because they exchange lists of their own graduates, the *CG* schools are less interested than the other institutions in either a computerized system or an open listing. A majority of the other institutions thought both of these recommendations would improve the market, while a majority of the *CG* schools disagreed.

TABLE 18—ATTITUDES OF ECONOMICS DEPARTMENT CHAIRMEN TOWARD OPEN LISTING AND COMPUTERIZED LISTING OF JOB CANDIDATES

(By Percent)

	CG Ph.D.	Other Ph.D.	Master's	Bachelor's
A computerized listing of job candidates would be useful in locating new faculty				
Agreeing	32.4	57.5	85.7	55.2
Disagreeing	50.0	25.0	0.0	27.6
Undecided	17.6	17.5	14.3	17.2
Open listing of employment opportunities would improve the functioning of the market				
Agreeing	35.3	47.5	64.3	58.6
Disagreeing	44.1	30.0	21.4	24.1
Undecided	20.6	22.5	14.3	17.2

X. Summary and Conclusions

A crisis of sorts has developed for new entrants in the labor market for economists—a crisis reflected to some extent in rising unemployment rates, but more significantly in growing underemployment. Choice academic jobs are becoming scarce relative to the supply of new Ph.D.s, thus forcing individuals who aspire to these jobs to accept positions with government, business and industry, and smaller colleges. In the past, job hunting has been a fairly easy task, with many graduates obtaining offers prior to entering the market. However, in the struggle to find employment in the market of the 1970's, many participants are at a loss as to what procedures to follow and are uncomfortable at the prospect of leaving the academic world for those of business and government. Their uncertainty is primarily a consequence of the informality of academic markets, which traditionally have been entered by two-thirds of all economists. Their discomfort results from the success criteria used by many economists, i.e., publications and faculty status at the more prestigious universities.

The burden of responsibility for finding employment should rest with the individual job seeker who, of course, stands to benefit most. However, the efficient allocation of manpower requires that decisions be informed decisions. Job seekers need to know what types of jobs economists hold, what functions they perform, what their salaries are, and how important that first job decision is relative to their futures. In addition, they need to know what techniques are most effective in locating employment in academia, business, and government.

Cross-sectional data are available from *NSF* on the distribution of economists by degree level, subfield, type of employer, work activity, annual salary, and geographic location. Information on placement techniques is less complete, but we do know that the academic market is the least formal. It functions most informally among the top-ranked departments, which hire primarily from each other. Even if these schools are required to list all vacancies with a central source, the market's complexion at this level is not likely to change. Those candidates who are known or whose sponsors are known are likely to be favored over those writing letters of application or responding to advertisements. Among the lower-ranked departments, such formal techniques as blind letters and registration with placement agencies are more successful. In business and government, direct inquiries are quite appropriate.

At this time not much is known about the extent of mobility within academia and between academia, business, and government. In the past, considerable mobility has been recorded, but it would be dangerous to assume that such patterns will hold over the next ten to fifteen years. Many observers, in fact, anticipate a decline in mobility.

Prospective economists should receive counseling early in their careers, aimed not at scaring them away from economics but at demonstrating that only a few will be able to achieve "professional bliss" as defined in the past. Those who decide to continue their education should be directed toward those fields of specialization which are in greatest demand. At least some departments may find it necessary to redesign their graduate programs, focusing more attention on teaching and application as opposed to research. At later stages, information can be supplied through lists of vacancies, frequent surveys of the market, and manuals describing job-seeking procedures in the profession.

* * * * *

Recommendations of the Committee on Hiring Practices

The fairly consistent evidence that recently graduated economists will have fewer job choices during the 1970's than was true during the 1960's has led to demands for improvements in the operation of the labor market for economists, improvements that would have been beneficial to all participants in this market even if the demand-supply ratio had not declined. The main concerns, especially of younger economists recently entering or about to enter the labor market, are that they will not have a fair chance at available jobs, either because of inadequate information or because of favoritism which causes the better jobs to go to a select few.

There is little the American Economic Association can do to change the informal system through which many jobs are filled, except to recommend that all of its members list their job openings and invite applications from a wider range of sources than the informal channels have ordinarily used. Efforts could be made to improve information about jobs in the government and business sectors as well as in academic institutions. Although open listing of positions would lead to more applications and therefore more screening work for employers, it also would widen the choices available to employers. Experiences in Great Britain, where academic positions are widely advertised but jobs continue to be filled informally, do not encourage optimism that open listing of jobs will do much to change prevailing informal labor market procedures. Nevertheless, we believe that measures can be taken in the vast majority of hiring situations which will improve the information and choices available to applicants and employers alike.

We therefore recommend that the American Economic Association:

(1) a. Collect and publish six times a year (February, April, June, August, October, and December) lists of job openings as well as lists of academic institutions with no openings. The cost of this service could be defrayed by subscription fees of, say, $20 per year. We should resolve that all members of the *AEA* have a professional obligation to list their openings with the association.

b. Conduct an annual labor market survey each September, to be published in December, perhaps with the December job list. This survey would include an estimate of the numbers of people expecting to enter the labor market and therefore provide some indication of the demand and supply relationship.

c. Explore with the U.S. Employment Service the possibility of establishing a cooperative arrangement to facilitate the distribution of labor market information for and about economists.

d. Establish cooperative arrangements with such organizations as the National Association of Business Economists, the American Bankers Association, and government employers of economists in order to acquire more information about the nonacademic demand for economists. These nonacademic sources become particularly important as academic job opportunities decline relative to the supply.

The publication of lists of jobs and applicants would make it possible to discontinue listing labor market information in the *American Economic Review*, thus freeing valuable space for other uses. There also are indications that very little use is made of the *AER* listings.

(2) Precede the annual meetings by one day for hiring transactions. There have been numerous complaints about confusion when department chairmen and others interested in interviewing candidates are attempting to do so while programs are in

progress. The labor market could be more orderly and those who wished to do so could attend more sessions if the first day were set aside for labor market transactions. Since programs would not be in session, better facilities for interviewing could be made available during the time reserved for hiring. Of course, hiring transactions would continue during the other days of the annual meetings, but hopefully at a less frantic pace than formerly if one day were devoted exclusively to this purpose.

It has sometimes been suggested that separate meetings be held for hiring purposes at some time during the year so as not to conflict with the annual meetings. However, this suggestion seems impractical, and setting aside a day during the annual meetings seems more appropriate.

(3) Make sure that adequate interviewing and communications facilities are available during the annual meetings. It would be particularly useful to provide effective communications systems for department chairmen or their representatives. The local arrangements committee should pay particular attention to planning the operation of labor market information facilities. If suitable arrangements cannot be made in the convention hotel, separate facilities should be established.

(4) Maintain a current list of department chairmen to be sent to participating departments and others on request. This would facilitate the communications process. Departments could send this information to the *AEA* along with news currently collected for announcements sections in the *AER*.

The American Economic Association could also encourage departments to take some actions to improve their placement and hiring procedures. These include:

(1) Recognizing a responsibility to help place students, including appointment of placement officers with adequate time and assistance to stay abreast of market developments.

(2) Counseling graduate students early about the present and probable future state of the labor market and how to behave in the labor market. Advisers should provide realistic information to students about their prospects.

(3) Providing full and accurate information to prospective employers. It is important to anticipate the kinds of data that would be useful to particular employers, with documentary evidence of teaching and research ability wherever available. It also is important not to describe the personal characteristics of women.

With respect to hiring, employers should:

(1) Wherever possible, make appointments with applicants in advance of the annual meetings.

(2) Exercise common courtesies in dealing with applicants. It is inexcusable for employers not to be open with applicants concerning their prospects, to interview several candidates at one time, not to give timely notification of the applicant's status, or to ask questions of women not asked of men.

While the *AEA* and graduate departments can do some things to facilitate a better matching of economists and jobs, major responsibility rests with *job applicants* to take the initiative and see that full and complete information is available to the appropriate people. Wherever possible, for example, applicants should encourage faculty members to supply letters of recommendation and other information to prospective employers.

The applicants should supply evidence of their scholarship and teaching ability wherever it is available and appropriate. They also should exercise such common courtesies as timely notification of acceptance of a position and openness in discussing their status. When writing to prospective employers, applicants should be encouraged to address department chairmen by name. Finally, when attending annual *AEA* meetings, applicants should make appointments in advance of the meetings and should carefully plan their schedules while there.

Members of the Committee on Hiring Practices are: F. Ray Marshall, chairman; Carolyn Shaw Bell, Robert J. Brown, Allan M. Cartter, and Stephen Marglin.[19]

APPENDIX A

Members of the Chairmen's Group (summer, 1973):[20]

Brown University
University of California, Berkeley
University of California, Davis
University of California, Los Angeles
Carnegie-Mellon Institute
University of Chicago
University of Colorado
Columbia University
Cornell University
Duke University
University of Florida
Harvard University
University of Illinois
Indiana University
Iowa State University
Johns Hopkins University
University of Maryland
Massachusetts Institute of Technology
University of Michigan
Michigan State University
University of Minnesota
New York University
State University of New York, Buffalo
University of North Carolina, Chapel Hill
Northwestern University
Ohio State University
University of Pennsylvania
University of Pittsburgh
Princeton University
Purdue University
University of Rochester
University of Southern California
Stanford University
Texas A&M University
University of Texas, Austin
Vanderbilt University
University of Virginia
University of Washington, Seattle
Washington State University
Washington University, St. Louis
Wayne State University
University of Wisconsin
Yale University

REFERENCES

C. S. Bell, "Women in Science: Definitions and Data for Economic Analysis," *Annals of the N.Y. Acad. of Sciences*, Mar. 15, 1973, *208*, 134–42.

F. M. Boddy, "Recent Behavior of Economists' Salaries," *Amer. Econ. Rev., Proc.*, May 1971, *61*, 316.

D. G. Brown, *The Market for College Teachers*, Chapel Hill 1965.

———, *The Mobile Professors*, Washington 1967.

A. M. Cartter, "Scientific Manpower for 1970–1985," *Science*, Apr. 9, 1971a, *172*, 132–40.

———, "Whither the Market for Academic Economists?" *Amer. Econ. Rev., Proc.*, May 1971b, *61*, 305–10.

M. F. Crowley, "Ph.D. Holders in Private Industry," *Monthly Labor Rev.*, Aug. 1970, *93*, 65–6.

R. B. Hitlin, "The Political Science Labor Market: A Report of the American Political Science Association Surveys and Activities," prepared for delivery at the 1973 annual meeting of the Amer. Pol. Sc. Assoc.

J. P. LaSalle, "Appointments, Promotion, and Tenure under Steady State Staffing," *Notices of the Amer. Math. Soc.*, Jan. 1972, *19*, 69.

M. C. Lovell, "The Production of Economic Literature: An Interpretation," *J. of Econ. Lit.*, Mar. 1973, *11*, 44.

H. D. Marshall, *The Mobility of College Faculties*, New York 1964.

J. R. Niland, "Allocation of Ph.D. Manpower

---

[19] At its December 1973 meetings, the Executive Committee of the American Economic Association approved these recommendations in principle, with the stipulation that the details of the job listing system were to be worked out before the March 1974 meeting of the Executive Committee.

[20] This group was later expanded from forty-three to sixty-three members.

in the Academic Labor Market," *Indust. Relat.*, May 1972, *11*, 141–56.

R. T. Prosser, "Note on the Job Market," *Notices of the Amer. Math. Soc.*, June 1970, *17*, 615.

B. E. Rhoades, "Preliminary Report of the *MAA* Committee to Facilitate Employer-Employee Contacts in Mathematics," *Amer. Math. Monthly*, Apr. 1972, *79*, 389.

G. G. Somers, "The Functioning of the Market for Economists," *Amer. Econ. Rev., Proc.*, May 1962, *52*, 509–10.

M. W. Stebbins, "The Social Science Ph.D. in Corporate Industry," Ph.D. diss., Univ. of Calif., 1972, distributed by the U.S. Dept. of Commerce, Natl. Tech. Inf. Serv., PB-213-533, 64.

G. J. Stigler, *Essays in the History of Economics*, Chicago 1965, 45.

R. P. Strauss, "A Younger Economist's Views on the Market," *Amer. Econ. Rev., Proc.*, May 1971, *61*, 327.

T. R. Vaughn and G. Sjoberg, "The Politics of Projection: A Critique of Cartter's Analysis," *Science*, July 14, 1972, *177*, 147.

G. S. Young, "The Problems of Employment in Mathematical Sciences," *Notices of the Amer. Math. Soc.*, Aug. 1971, *18*, 718–22.

American Council on Education, *A Fact Book on Higher Education*, first issue 1973.

American Historical Association, *Newsletter*, Sept. 1972, 4.

American Mathematical Society, *Amer. Math. Monthly*, Jan. 1973, *80*, 115.

American Political Science Association, "*APSA* Council Minutes," *Polit. Sc.*, Winter 1971, *4*, 73.

Business Week, "Corporate Economists Carve a New Domain," *Bus. Wk.*, Nov. 28, 1970, 74.

Minutes of the Annual Meeting of the American Economic Association, Dec. 28, 1971, *Amer. Econ. Rev., Proc.*, May 1972, *62*, 474.

Modern Language Association, "*MLA* Job Information Service Subscriber Survey," *MLA Newsletter*, Sept. 1972, *4*, 3.

National Science Foundation, *American Science Manpower, 1970*, Washington 1972a.

———, *Unemployment Rates and Employment Characteristics for Scientists and Engineers, 1971*, Washington 1972b.

Report of the Secretary for 1972, *Amer. Econ. Rev., Proc.*, May 1973, *63*, 480.

U.S. Dept. of Commerce, Bur. of the Census, "Detailed Characteristics," *United States Census of Population: 1970*, final report PC (1)-D1, Washington 1973, 718.

U.S. Dept. of Labor, Bur. of Labor Statist., *Occupational Outlook Handbook 1972–1973*, bull. no. 1700, Washington 1972, 202.

U.S. Office of Education, *Earned Degrees Conferred, 1969–1970*, Washington 1972.

REPORT OF COMMITTEE ON ECONOMIC EDUCATION

In 1972, this Committee laid out an agenda for improving the teaching of economics which was published in the *American Economic Review*, in May 1973 (pp. 303–308). During 1973, we pushed forward on the major fronts outlined in that agenda. Specifically, we report four main activities.

(1) *Improved training for new teachers.* The major undertaking in our 1972 agenda was to help university departments provide improved training in teaching for new Ph.D.s. The program outlined in the agenda has been implemented, with the help of a grant of approximately a quarter of a million dollars from the Sloan Foundation. In August 1973, a pilot-plant program for young teachers just completing their Ph.D.s at several large universities was conducted at the University of Indiana under the auspices of this Committee and the Joint Council on Economic Education. The goal was to help such potential teachers on both the techniques of teaching and the subject matter of basic economic courses which they would probably be asked to teach. Greater detail is provided by the agenda offered by the Committee in the May 1973 issue of this *Review*. About thirty-five new Ph.D.s, two or more from each institution represented, plus senior professors from most of the same institutions, participated in the pilot-plant seminar.

In autumn 1973, teacher-training programs, based on the pilot-plant operation at Indiana, were inaugurated at the Universities of Minnesota and Wisconsin. Additional schools will be added over the next two or three years, reaching a total of at least ten within the next five years. One output of the pilot-plant program in 1973 is a series of suggestions and video tapes, plus a "manual," which we hope will be useful to various universities throughout the country that want to do more to help their new Ph.D.s become prepared to be effective economics teachers. More universities have already indicated their interest in participating than were planned for the undertaking.

(2) *Subject matter help for teachers.* In its agenda, the Committee recommended that the profession do more to provide subject matter help for teachers of economics, especially those at community and junior colleges and at other less known colleges and universities where teachers often find it difficult to keep up with new theory and policy developments and to do an adequate job of teaching a variety of courses. We urged that the *Journal of Economic Literature* adopt a policy of devoting half of its survey articles to helping such teachers keep up with developments in theory and policy. We also recommended a new series of *Readings* volumes, probably to be published annually, under the sponsorship of this Committee and the Association, including articles with the same goal as that indicated just above. The editor of the *Journal of Economic Literature* and the Executive Committee of the Association have both expressed sympathy with the goal of including such articles in the *Journal of Economic Literature.* Moreover, arrangements have been completed to begin the new series of paperback *Readings* volumes, to be published annually by Richard Irwin, Inc., under the editorship of Rendigs Fels and John Siegfried of Vanderbilt University and the general policy guidance of this Committee. It will include a series of articles each year, re-

printed from economic journals and other places, aimed at pulling together in one place such updating materials. Contents of the first volume have been chosen, and we are encouraged by the editors' success in assembling the first volume.

(3) *Better information on teaching experiments and experience.* Many experiments and much experience with new developments in teaching economics are lost to the profession because there has been no adequate basis for sharing them with other teachers who may be interested. The *Journal of Economic Education*, established a few years ago at the suggestion of this Committee and published by the Joint Council on Economic Education, issued a special volume in autumn 1973 on "The Evaluation of Effectiveness in Teaching Economics," which we recommend to the profession. Moreover, this *Journal* is now providing in its regular issues the rapidly growing body of experience on pilot-plant experiments of various sorts (computer-aided instruction, programmed learning, games, and the like) in teaching economics, plus reports on especially devised elementary courses at different institutions with the goal of aiding those teachers who do not have the resources to develop new course materials of their own.

(4) *Research on the efficiency of economic education.* This Committee has continued to encourage research by teachers of economics on the efficiency of different approaches and methods in teaching college economics. We are happy to report a rapidly increasing volume of such research, much of which is being reported through the *Journal of Economic Education* and other economic journals. We believe that such research, shared with other teachers and subjected to replication and criticism, can provide an important foundation for improving the teaching of undergraduate economics throughout the country.

The Committee invites the comments, suggestions, and criticisms of members of the profession on its activities. We especially urge that teachers experimenting with new approaches and methods in undergraduate courses, especially where the results are carefully evaluated, make these results known to others by submitting articles to the *Journal of Economic Education* or to other appropriate journals.

G. L. BACH, *Chairman*

REPORT OF THE CENSUS ADVISORY COMMITTEE

The Association's Advisory Committee to the U.S. Bureau of the Census, composed of fifteen members, met twice during 1973. At its April 2–3 meeting, the agenda included a review of the Bureau's future program plans and the following topics: staff development; Agricultural Census plans for 1974 and 1977; 1972 Economic Censuses—progress report and publication plans; the Demographic Periodic Program—Decennial Census of Population and Housing and a proposed 1975 Survey of Social and Economic Characteristics of Population; a review of activities in the household survey area; dissemination of Census data including microdata samples; Survey of Manufacturing Capacity—review of preliminary results; seasonal adjustment of weekly retail sales; and a general discussion including the role of the Committee.

The October 11–12 meeting began with an evening session at the National Economists Club. Committee members were greeted by Sidney L. Jones, Assistant Secretary of Commerce for Economic Affairs, and Vincent P. Barabba, the new Director of the Bureau. This session mainly was devoted to a response to the Committee's recommendations arising from the April 1973 meeting. The following day agenda topics included: confidentiality and disclosure policy; preliminary thoughts on the 1980 Census of Population; the environmental statistics program, including pollution abatement; wage rate data in the Current Population Survey (*CPS*); reconciliation (between countries) of international trade statistics; compatibility and reliability of government statistical series; and a general discussion.

The Committee has made many detailed recommendations to the Bureau in the above topic areas (space limitations preclude reporting those here, but copies of minutes of the meetings are available through the Director's office). The principal concern of the Committee has been preserving and increasing the strength and vitality of the Bureau's statistical programs.

In the past few years, the question of the need for data collected by the Bureau and the confidentiality of Census information, protected by Title 13 of the U.S. Code, have been under attack by various members of the Congress and special interest groups. The Bureau feels strongly, and the Committee agrees, that the high sample response rate and quality of U.S. Census data in large measure are due to legally guaranteed assurances of confidentiality; only Census personnel or sworn agents with a need to know have access to individual or company Census records. While recognizing the need for information for scholarly and scientific purposes, it would be well for the Association, by resolution of the membership, to reaffirm its support for the U.S. Bureau of the Census and the need for continued appropriate protection of confidentiality of past, present, and future information on individual persons, families, or firms collected by the Bureau.

Another major area of concern of the Committee within recent years has been the appointment of Committee members, the role and effectiveness of the Committee in advising the Bureau, and its responsibilities and functions in informing the Association's membership and mobilizing support for vital statistical programs.

It is recommended that the Association

appoint an *ad hoc* committee to review these matters, to draft recommendations and resolutions, and to report to the Executive Committee at its spring 1974 meeting.

Five members, including the present chairman, are retiring from the Committee this year: Charles C. Holt, John W. Kendrick, Murray Weidenbaum, and Richard B. Wirthlin. The Chairman-elect and Vice-chairman-elect are, respectively, Lee E. Preston and Richard Ruggles.

GARY FROMM, *Chairman*

REPORT OF REPRESENTATIVE TO THE INTERNATIONAL ECONOMIC ASSOCIATION

During the calendar year 1973 the International Economic Association held three conferences and one workshop in four countries on two continents: Bangladesh, Japan, Western Germany, and France.

The conference in Dacca, Bangladesh, January 7 to 12, on "Economic Development of Bangladesh Within the Framework of a Socialist Economy," was prepared by a program committee consisting of Austin Robinson (United Kingdom), Nurul Islam (Bangladesh), Bela Czikos-Nagy (Hungary), with Swadesh Bose as a local secretary. Three economists from the United States participated: Hollis Chenery, Gus Ranis, and Jaroslav Vanek.

In Tokyo, Japan, a conference on "The Economics of Health and Medical Care" was held from April 2 to 7. Shigeto Tsuru (Japan) and Mark Perlman (United States) were co-chairmen of the program committee, which included Martin S. Feldstein (United States) and Hirofumi Uzawa (Japan). Professor Lisycin (USSR) was invited too but, regrettably, was not able to serve. The following economists from the United States participated: Karen Davis, Bernard Friedman, Victor R. Fuchs, Michael Grossman, Michael Intriligator, Barbara H. Kehrer, Judith Lave, Lester Lave, Samuel Leinhardt, Fritz Machlup, Mark Perlman, Charles E. Phelps, Richard N. Rosett, and Frank A. Sloan.

A workshop on Mathematical Economics was held at Rheda Castle (University of Bielefeld), Western Germany, from June 17 to July 14, under the direction of Christian von Weizsäcker, assisted by Professors Hildenbrand, Schönfeld, and Selten. The workshop was attended by twenty-seven economists from twelve countries, none from the United States.

In Valescure, France, a conference on "Economic Aspects of Population Growth" was held from September 3 to 8. The program committee consisted of Ansley J. Coale (United States), Chairman, Milos Macura (Yugoslavia), Marc Nerlove (United States), N. V. Sovani (India), Leon Tabah (France), and E. A. Wrigley (United Kingdom). The following economists from the United States attended as participants or observers: Bryan Boulier, Ansley J. Coale, Paul Demeny, Albert I. Hermalin, I. M. Hume, Laurence Kegan, Ron Lesthaeghe, Fritz Machlup, Mark Perlman, Ronald G. Ridker, T. Paul Schultz, Frederic C. Shorter, and Etienne Van de Walle. Michael P. Todaro contributed a paper but could not attend.

In 1974 the Association will hold one conference, one workshop, and its fourth World Congress. The conference, on "The Economics of Public Services," will be at the Agnelli Foundation in Turin, Italy, from April 1 to 6. Martin S. Feldstein is chairman of the program committee. The World Congress, on "Economic Integration: Worldwide, Regional, Sectoral," will be held in Budapest, Hungary, from August 19 to 24. The program committee, consisting of twelve economists from eleven countries, is chaired by Fritz Machlup. The program will be published in the Notes of the March and June issues of the *American Economic Review.* At the time of this report, the program contains the names of sixty-seven economists from thirty-seven different countries on all continents, but it provides for the participation of hundreds of others in ten working groups, each devoted to a specific aspect of international economic integration. A workshop on "Population and Development," designed for officials and economists from developing countries, will be held at Bosphorus University, Turkey, under the direction of Paul Demeny, assisted by Frederic C. Shorter and Imre Gönensay.

Plans for 1975 include a spring conference, probably in Spain, on "The Microeconomic Foundations of Macroeconomic Theories." The program committee is chaired by Geoffrey Harcourt (Australia), assisted by Sir John Hicks and four others. A program committee under the chairmanship of Erik Lundberg (Sweden) will be constituted for an autumn conference on "Inflation."

Five volumes of proceedings were published during 1973: *Latin America in the International Economy*, edited by Victor L. Urquidi and Rosemary Thorp; *Science and Technology in Economic Growth*, edited by Bruce P. Williams; *Models of Economic Growth*, edited by James A. Mirrlees and N. H. Stern; *Allocation under Uncertainty*, edited by Jacques H. Dreze; and *Transport and the Urban Environment*, edited by Jerome Rothenberg and T. G. Heggie. Two other volumes are at present with the printer and another two are being prepared for publication.

FRITZ MACHLUP

REPORT OF THE REPRESENTATIVE TO THE NATIONAL BUREAU OF ECONOMIC RESEARCH

The most important development of the Bureau's year was the establishment of a West Coast office in Stanford, California. It is expected to begin operation in early 1974 under the supervision of Co-directors Victor R. Fuchs, and Sherman J. Maisel. The office will be a research center through which the Bureau hopes to establish closer relations with the academic and business communities of the area and to expand its exposure to national and international problems. In addition to the headquarters in New York City, the Bureau already has branch offices in Cambridge and New Haven.

The Bureau's research program involved more than a hundred individual projects during the year. They and a listing of the Bureau's present and prospective publications are described in the 53rd Annual Report, available on request from the Bureau, 261 Madison Ave., New York, N. Y. 10016.

Conferences were held in 1973 on "Household Production and Consumption," "Economics of the Environment," "Research and the 1970 Public Use Sample," and "Consumer Prices and Expenditures." In addition, there were a number of on-going working groups, centered at various host universities. Future conferences being organized will deal with "Distribution of Economic Well-being," "Price Behavior," "Urban Spatial Structure," "Role of Health Insurance," "Economic Analysis of Political Behavior," "National and Regional Accounts," and "Microdata for Economic and Social Research."

At the Annual Meeting on September 17, Philip M. Klutznick became the Director representing the Committee for Economic Development; Roy E. Moore, representing the National Association of Business Economists; William S. Vickrey, representing Columbia University; and Emilio G. Collado, a Director-at-large. Frank W. Fetter was elected a Director Emeritus, having served as a Director since 1950. The top officers were reelected, namely, Walter W. Heller, Chairman, J. Wilson Newman, Vice-Chairman, and John R. Meyer, President.

WILLARD L. THORP

REPORT OF THE COMMITTEE ON THE STATUS OF WOMEN IN THE ECONOMICS PROFESSION

During the past year the Committee has widened the market for economists, increased the effective supply of women economists, served as a model for the efforts of other professional associations, and added to information about the economics profession.

1. Over one thousand women economists have received four Newsletters during 1973. They list not only employment opportunities but requests for articles, conference planners and participants, referees and reviewers, etc. Over the past year women economists have appeared on more programs, advisory boards, and the like than previously; the Ford Foundation, typical of nonacademic institutions, reports that many more women have applied for their openings; at the December 1973 meetings of the American Economic Association (*AEA*), women were interviewed in large numbers for positions they had previously applied for. The Committee does not take full credit for these happenings and other evidence that women economists are beginning to enter the market of ideas and of employment; nevertheless, we think the Newsletter has paid off. The impact of our publicity has been to widen the market for all economists for two reasons. First, the Committee has not restricted its information to women (male graduate students were grateful to obtain at the meetings copies of our job listings). Second, the Committee reports *all* opportunities, whether these come to us directly or indirectly. From Ohio State University we hear, "I think that the attempt to fulfill employment agency functions is commendable. My experiences with the labor market for economists during the past five years suggests that the profession is too bound by tradition and cronyism. The efforts of the Committee in this area may well help break up some of the invidious segmentation which characterizes this labor market."

2. The growing roster compiled by the Committee is our mailing list for the women economists receiving the Newsletter. It currently numbers about 1400, compared to about 300 in early 1973. In effect, we have increased the supply of women economists by finding them. The roster has been computerized, following authorization by the Executive Committee, to provide information, for each woman listed, on her field, education, experience, availability, etc. At its March 1973 meeting, the Executive Committee agreed to operate the roster for employment purposes once it had been set up. At its November meeting the Committee decided to provide lists to prospective employers and others of women economists, sorted by field. One inquirer for such a list remarked, "It will be very useful to have a roster such as yours. . . . everyone knows a few women economists, but nobody knows many—and your list . . . is a great improvement on my colleagues' recollection of ex-fellow students." At its December meeting, the Committee decided that it should not provide this service on a zero-price basis, and anyone requesting such a list in the future will be asked to help defray costs.

3. The survey of academic economists conducted in the fall of 1972 has been widely used by the Committee and others. Some findings were quoted in the Carnegie Commission report, *Opportunities for Women in Higher Education* (which also commended this Committee's work on

TABLE 1—DEPARTMENTS OF ECONOMICS SURVEY RESULTS BY SIZE[a]

Number of Students	Questionnaires Sent	Questionnaires Returned	Response Rate (In Percent)
1– 499	145	3	2
500– 999	333	53	16
1,000– 1,499	238	53	22
1,500– 1,999	122	38	31
2,000– 2,999	117	40	34
3,000– 4,999	143	49	34
5,000– 9,999	165	78	47
10,000–14,999	68	42	62
15,000–24,999	56	38	68
25,000 or more	22	20	91
TOTAL	1,409	414	29
10,000 or more	146	100	68

[a] Size refers to the number of students, both graduate and undergraduate.

opening up the job market). Our figures on the percentage of women enrolled for higher degrees have enabled many institutions to set up goals for affirmative action programs and have encouraged others to recruit more widely for faculty and graduate school applicants. There is a clear need for another follow-up survey, given this year's results which are summarized below.

4. The findings, recommendations, and conclusions of the Committee, with respect to implementing the resolutions adopted by the Association at its December 1972 business meeting, appeared in the December 1973 *American Economic Review*. We hope that the guidelines set forth in the article will be widely used in academic and nonacademic organizations. Entitled "Combatting Role Prejudice and Sex Discrimination," this statement was discussed at the session sponsored by the Committee on December 28, 1973. We are grateful to George Stigler, David Gordon, and Elizabeth Clayton for reviewing the article at that time. Reprints of the article (for our cost of $2.50) and tapes of the session (for our cost of $15) can be purchased from the Executive Secretary, Helen Bolln Munzer, 131 Kent Street, Brookline, Mass. 02146.

5. The Committee has served as a model for other professional associations which have sought help from us on a variety of problems. The Executive Secretary has provided technical advice about surveys, statistical tabulations, compiling and operating a roster, cataloging employment opportunities, making referrals, and so on. The Chair has talked with many people, most of whom were referred to us by others, about such committees and about the role of professional associations in dealing with questions of discrimination. We take these inquiries as recognition of our accomplishments; we also note with pleasure the overwhelming support from women economists in general. Almost all professional associations have set up some program on behalf of women, and we have been in touch with most committees, commissions, and the like. In many professional associations, a women's caucus or other dissident group of women members has appeared, presumably because the formal efforts seem obviously *pro forma*. That such a group has not appeared within the *AEA* may be an endorsement of our attempts to deliberately plan for effective action and to provide a normal channel for the expression of legitimate grievances.

5. Typical of the ongoing efforts of the Committee have been the letters sent to presidents and program chairmen of regional economic associations and other allied associations, as well as to editors of economic journals, urging their appointment of women to policy positions. We have also urged the editor of the *Journal of Economic Literature*, who has agreed, to commission a survey article on the literature about the economic status of women. The Committee has a close liaison with the Federation of Organizations for Professional Women, which may succeed in centralizing the efforts of a good many pro-

TABLE 2—NONRESPONSE TO THE QUESTIONNAIRE

Departments of Economics With More Than 10,000 Students

University of Akron[1]
Arizona State University
Brigham Young University[1]
California State College–Los Angeles
California State University–Northridge
Central Michigan University
Central Missouri State College[1]
Central State College–Oklahoma[1]
University of Connecticut–Storrs
Eastern Michigan University
Florida State University
Fordham University
George Washington University[1]
Georgia State University
University of Georgia
Hofstra University
Illinois State University
University of Illinois–Chicago Circle
Long Island University–Greenvale[1]
Long Island University, C. W. Post Center[1]
University of Maine–Bangor
Mankato State College[1]
Marquette University[1]
Memphis State University[1]
University of Miami
University of Nebraska–Omaha
City University of New York
CUNY–City College
CUNY–Queens College
Newark State College[1]
Ohio University
Oklahoma State University
Old Dominion University[1]
University of Puerto Rico[1]
Rochester Institute of Technology[1]
San Francisco State College[1]
University of South Florida[1]
University of South Carolina
Temple University[1]
University of Tennessee
University of Texas–Arlington[1]
University of Utah
Virginia Commonwealth University[1]
Virginia Polytechnic Institute
Walla Walla College[1]
Wisconsin State University–Oshkosh

[1] Also did not return last year's questionnaire.

fessional associations, as well as looser ties with Women's Equity Action League, American Association for the Advancement of Science, etc. etc. etc.

7. We have received only three requests for direct intervention on behalf of women economists—two involving individuals and the other a textbook. In the latter case, we pointed out offensive passages and style to the author, who has assured us of substantive revision. Guidelines for writers, especially of elementary texts, should clearly be on our agenda. As for the other two cases, one individual requested that we take no action, after we had indicated to her the avenues of redress available through American Association of University Professors (*AAUP*), Equal Employment Opportunities Commission (*EEOC*), and Office of Contract Compliance and had offered to ask for explanation from the officials involved. The other case has not yet been settled. The Committee's function has been chiefly to inform the parties concerned of the resolutions of the *AEA* about discrimination, of the remedies available through *AAUP*, *EEOC*, and the Wages and Hours Division of the Department of Labor. It is also true that this particular complainant has received significant upward adjustment in rank, salary, and privileges. The Committee believes firmly that the Association can be most helpful not by acting as *amicus curiae* in a court suit but by emphasizing the professional responsibility to cooperate fully with both law and procedure. We also believe that the threat of adverse publicity remains a strong inducement to compliance.

7. Accordingly, we present some summary figures on the employment of women economists on faculties of universities and colleges and on the enrollment of women students. The figures are not directly comparable to those of last year because

TABLE 3—STUDENTS AND FACULTY IN ECONOMICS

Full-time and Part-time Totals, Women as Percent of Totals, January 10, 1974

Faculty	All Departments				New Cartel				Old Cartel				Noncartel			
	Full-time (Total)	(Percent)	Part-time (Total)	(Percent)	Full-time (Total)	(Percent)	Part-time (Total)	(Percent)	Full-time (Total)	(Percent)	Part-time (Total)	(Percent)	Full-time (Total)	(Percent)	Part-time (Total)	(Percent)
All Ranks	4,262	7	574	13	1,533	5	185	12	1,224	5	152	14	2,729	8	389	14
Professor	1,439	3	70	9	714	2	39	5	592	2	32	6	725	4	31	13
Associate	1,074	6	34	9	325	3	7	0	239	3	7	0	749	7	27	11
Assistant	1,371	9	64	19	430	9	29	24	357	9	23	26	941	9	35	14
Instructors	292	14	116	17	36	17	19	5	12	17	14	7	256	14	97	20
Lecturers	86	16	290	12	28	29	91	14	24	29	76	16	58	10	199	11
Staff (not included in totals)	100	94	66	58	67	95	29	55	55	96	19	74	33	91	37	59
Students	**All Departments**				**New Cartel**				**Old Cartel**				**Noncartel**			
	Full-time (Total)	(Percent)	Part-time (Total)	(Percent)	Full-time (Total)	(Percent)	Part-time (Total)	(Percent)	Full-time (Total)	(Percent)	Part-time (Total)	(Percent)	Full-time (Total)	(Percent)	Part-time (Total)	(Percent)
Senior Majors	9,423	15			3,065	11			1,951	12			6,358	17		
Ph.D. Candidates	4,379	12	648	9	3,631	12	419	9	3,240	12	268	9	748	13	229	9
M.A. Candidates	2,724	14	1,585	15	1,199	12	703	13	860	12	209	17	1,525	15	882	16
Scholarships	563	17			455	12			434	11			108	35		
Fellowships	1,009	16			803	13			704	14			206	27		
Assistantships	2,294	15			1,588	13			1,327	13			706	19		

the respondents differ. The "major institutions" cited have also changed. The so-called Chairman's Group identified by the *AEA* consisted last year of forty-three *U.S.* universities which graduated two-thirds of the Ph.D.s in economics plus one foreign university. This year the *AEA* enlarged this group. Our figures refer to fifty-seven *U.S.* universities, which presumably grant most of the Ph.D. degrees in economics. The new Chairman's Group also contains two foreign universities and eight nonuniversities, most of which are non-profit institutions. Obviously we wanted to compare last year to this year; just as obviously we want to distinguish between major institutions (even if "major" is defined differently) and others. The tabulations refer, therefore, to different groups of so-called major universities; the groups were not determined by this Committee, but by the *AEA*. We have used the term "Old Cartel" to refer to the 1972 list of institutions and "New Cartel" for the 1973 list.

As students at the graduate level, women account for the same fraction of full-time Ph.D.s (12 percent) as last year; part-time Ph.D.s were more numerous among the men enrolled than among the women (9 percent of the total). Twelve percent of the candidates for M.A.s were women, but at this level women accounted for a larger number of part-time students (15 percent at all universities, 17 percent at the major schools). Figures on scholarships, fellowships, and assistantships show that the proportion going to women was about the same as the proportion of women in the student body. The most significant finding about women students of economics deals with undergraduate majors, where there is a sharp distinction between the cartel schools and the rest. At the major universities, women account for only 11 percent of the senior majors in economics; at the noncartel schools, they make up 17 percent of the senior majors.

Evidence from other disciplines suggests that the number of women students in a given field is strongly correlated with the number of women faculty; data from this survey lend support to this hypothesis. Eight percent of the faculty at noncartel schools are women, compared to only 5 percent at the cartel schools. On the other hand, all-male faculties exist at 21 percent of the "old cartel" schools, 30 percent of the "new cartel" schools, and 60 percent of the noncartel institutions. Consequently, the 212 women faculty members outside the walls of the major universities are concentrated (if that is the right word) in only 142 schools, or 74 percent of the women economists are employed by 40 percent of the institutions. Prominent among these, of course, are women's colleges where both faculty representation and student enrollment in economics are impressively high.

One year is obviously too short a time to draw any conclusions about change. The only comparable data base consists of the forty-three universities termed the "old cartel." Last year eighteen of these institutions had no women faculty in economics; this year ten remain totally male. As a group, the old cartel increased its women faculty by fourteen percent (nine women). The net increase, however, resulted from hiring twenty women to these economics faculties.

Based on 414 respondents with 4,262 full-time economics faculty members, women accounted for 7 percent of the total, compared to 6 percent a year ago; this represented, however, 76 out of 1,533 faculty at the new cartel schools and 212 out of 2,729 everywhere else. For all respondents, women were 3 percent of the professors, 6 percent of the associate, and 9 percent of the assistant professors, 14 percent of the instructors, and 16 percent of the special lecturers, with more women than last year holding positions on the tenure track. Women gained 8 percent of all promotions (20 of 250), although at the new cartel universities precisely three women were granted higher rank, one each as professor, associate, and assistant (3 of 74). These same universities hired twenty-four women faculty, but their net gain over last year was only nine women. Other questions about the new cartel schools revealed that about two-thirds have special procedures for hiring women, although most of them do not give entry-level appointments to their own Ph.D.s. Almost all have nepotism rules parallel to those of the *AAUP*, slightly under half allow husband and wife to share an appointment, and more than two-thirds of these institutions have reviewed their salaries for evidence of discrimination. As for part-time faculty, one-third of the institutions grant fringe benefits and a smaller number consider part-time faculty for tenure. On the other hand, over half have part-time faculty participating in department meetings. As for increasing their own supply of female students, about 25 percent of these major universities actively recruit women; about 75 percent have time limits within which the Ph.D. must be earned, but only one has an age limit for students.

CAROLYN SHAW BELL, *Chair*

Xerox Individualized Publishing is good for the environment. It eliminates waste.

What is Xerox Individualized Publishing (XIP)? How does it eliminate waste? XIP lets you assemble a custom anthology with the help of our exclusive catalog and trained editorial staff. Since you create exactly the Reader/text you want, **students need only buy one book.** Since there is **no restriction on the amount of original material you may include**, you can stop running to the duplicating machine and put your course outline, term paper topics, introcuctory comments, or your own previously unpublished work between one set of covers. **Our catalog lists and describes the best popular and scholarly articles and book excerpts in your field.** Avoid the stacks and edit your XIP Reader/text in the comfort of your own home while our staff does the legwork. And since **we secure permissions** for each article, XIP Reader/text is completely legal. That's how you liminate wasted time, wasted energy, wasted money. And, you save a lot of trees.

So do your bit for the environment. Become a XIP editor today. Fill in this coupon and mail it to us: Xerox College Publishing, 191 Spring Street, Lexington, Mass. 02173.

.... Please send me a XIP catalog

.... Please have a sales representative call on me.

Name ....

School ....

Address ....

City ....

State .... Zip .... Phone ....

Enrollment/year ....

AmER

Xerox College Publishing
191 Spring Street
Lexington, Massachusetts 02173

General Economic Equilibrium: Purpose, Analytic Techniques, Collective Choice

By Kenneth J. Arrow*

I. The Coordination and Efficiency of the Economic System

From the time of Adam Smith's *Wealth of Nations* in 1776, one recurrent theme of economic analysis has been the remarkable degree of coherence among the vast numbers of individual and seemingly separate decisions about the buying and selling of commodities. In everyday, normal experience, there is something of a balance between the amounts of goods and services that some individuals want to supply and the amounts that other, different individuals want to sell. Would-be buyers ordinarily count correctly on being able to carry out their intentions, and would-be sellers do not ordinarily find themselves producing great amounts of goods that they cannot sell. This experience of balance is indeed so widespread that it raises no intellectual disquiet among laymen; they take it so much for granted that they are not disposed to understand the mechanism by which it occurs. The paradoxical result is that they have no idea of the system's strength and are unwilling to trust it in any considerable departure from normal conditions. This reaction is most conspicuous in wartime situations with radical shifts in demand. It is taken for granted these can be met only by price control, rationing, and direct allocation of resources. Yet there is no reason to believe that the same forces that work in peacetime would not produce a working system in time of war or other considerable shifts in demand. (There are undesirable consequences of a free market system, but sheer unworkability is not one of them.)

I do not want to overstate the case. The balancing of supply and demand is far from perfect. Most conspicuously, the history of the capitalist system has been marked by recurring periods in which the supply of available labor and of productive equipment available for the production of goods has been in excess of their utilization, sometimes, as in the 1930's, by very considerable magnitudes. Further, the relative balance of overall supply and demand in the postwar period in the United States and Europe is in good measure the result of deliberate governmental policies, not an automatic tendency of the market to balance.

Nevertheless, when all due allowances are made, the coherence of individual economic decisions is remarkable. As incomes rise and demands shift, for example, from food to clothing and housing, the labor force and productive facilities follow suit. Similarly, and even more surprising to the layman, there is a mutual interaction between shifts in technology and the allocation of the labor force. As technology improves exogenously, through innovations, the labor made redundant does not become permanently unemployed but finds its place in the economy. It is truly amazing that the lessons of both theory and over a century of history are still so misunderstood. On the other hand, a growing accumulation of instruments of production raises real wages and in turn induces a rise in the prices of labor-intensive commodities relative to those which use little labor. All these phenomena show that by and large and in the long view of history, the economic system adjusts with a considerable degree of smoothness and indeed of rationality to changes in the fundamental facts within which it operates.

---

* Harvard University. This article is the lecture Kenneth Arrow delivered in Stockholm, Sweden, December 1972, when he received the Nobel Prize in Economic Science. The article is copyright © the Nobel Foundation 1973. It is published here with the permission of the Nobel Foundation, and is included in the volume of *Les Prix Nobel en 1972*.

The problematic nature of economic coordination is most obvious in a free enterprise economy but might seem of lesser moment in a socialist or planned society. But a little reflection on the production and consumption decisions of such a society, at least in the modern world of complex production, shows that in the most basic aspects the problem of coordination is not removed by the transition to socialism or to any other form of planning. In the pure model of a free enterprise world, an individual, whether consumer or producer, is the locus both of interests or tastes and of information. Each individual has his own desires, which he is expected to pursue within the constraints imposed by the economic mechanism; but in addition he is supposed to have more information about himself or at least about a particular sphere of productive and consumptive activity than other individuals. It might be that in an ideal socialist economy, all individuals will act in accord with some agreed ideas of the common good, though I personally find this concept neither realistic nor desirable, in that it denies the fact and value of individual diversity. But not even the most ideal socialist society will obviate the diversity of information about productive methods that must obtain simply because the acquisition of information is costly. Hence, the need for coordination, for some means of seeing that plans of diverse agents have balanced totals, remains.

How this coordination takes place has been a central preoccupation of economic theory since Adam Smith and received a reasonably clear answer in the 1870's wit the work of Jevons, Menger, and above all Léon Walras: it was the fact that all agent in the economy faced the same set o prices that provided the common flow o information needed to coordinate the system. There was, so it was argued, a set of prices, one for each commodity, which would equate supply and demand for all commodities; and if supply and demand were unequal anywhere, at least some prices would change, while none would change in the opposite case. Because of the last characteristics, the balancing of supply and demand under these conditions may be referred to as *equilibrium* in accordance with the usual use of that term in science and mathematics. The adjective, "general," refers to the argument that we cannot legitimately speak of equilibrium with respect to any one commodity; since supply and demand on any one market depends on the prices of other commodities, the overall equilibrium of the economy cannot be decomposed into separate equilibria for individual commodities.

Now even in the most strictly neoclassical version of price theory, it is not precisely true that prices alone are adequate information to the individual agents for the achievement of equilibrium, a point that will be developed later. One brand of criticism has put more stress on quantities themselves as signals, including no less an authority than the great Keynes (1936); see especially the interpretation of Keynes by Leijonhufvud (1968, especially ch. 2). More recently the same argument has been advanced by Kornai (1971) from socialist experience. Nevertheless, while the criticisms are, in my judgment, not without some validity, they have not given rise to a genuine alternative model of detailed resource allocation. The fundamental question remains, how does an overall total quantity, say demand, as in the Keynesian model, get transformed into a set of signals and incentives for individual sellers?

If one shifts perspective from description to design of economies it is not so hard to think of nonprice coordinating mechanisms; we are in fact all familiar with rationing in one form or another. Here, the discussion of coordination shades off in that of efficiency. There has long been a view that the competitive price equilibrium is efficient or optimal in some sense that rationing is not. This sense and the exact statement of the optimality theorem were clarified by Pareto (1909, ch. 6, sections 32–38) and, in the 1930's by my teacher, Harold Hotelling (1938) and by Abram Bergson (1938). An allocation of resources is *Pareto efficient* (or *Pareto optimal*) if there is no other feasible allocation which will make everyone better off (or, as more usually stated, make everyone at least as well off and at least one member better off). Then, by an argument that I shall sketch shortly, it was held that a competitive equilibrium necessarily yielded a Pareto-efficient allocation of resources.

It was, of course, recognized, most explicitly perhaps by Bergson, that Pareto efficiency in no way implied distributive justice. An allocation of resources could be efficient in a Pareto sense and yet yield enormous riches to some and dire poverty to others.

II. The Hicks-Samuelson Model of General Equilibrium

I will state more formally the model of general competitive equilibrium as it had been developed by about 1945, primarily through the detailed developments and syntheses of John Hicks (1939) and Paul Samuelson (1947). Competitive analysis is founded on two basic principles: optimizing behavior on the part of individual agents in the presence of prices taken as given by them and the setting of the prices so that, given this individual behavior, supply equals demand on each market. The outcome of the competitive process is then to be evaluated in terms of Pareto efficiency and additional conditions on the resulting distribution of goods.

The maximizing behavior of individuals has been well surveyed by Samuelson in his Nobel lecture (1971), and I will not go over that ground here. I just want to remind the listener of a few elementary points. The first is that the consumer's choices are subject to a budget constraint. The consumer starts with the possession of some quantities of economically valuable goods, such as labor of particular types, land, or other possessions. Let us imagine there are n commodities altogether, and let $\bar{x}_{hi}$ be the amount of commodity i owned initially by individual h (this may well be zero for most commodities). If p_i is the price of the ith commodity, then his total income available for expenditure is

$$\sum_{i=1}^{n} p_i \bar{x}_{hi}$$

Hence, he can choose for consumption any bundle of goods, $x_{h1}, \ldots, x_{hn}$, which cost no more than his income,

$$\sum_{i=1}^{n} p_i x_{hi} \leqq \sum_{i=1}^{n} p_i \bar{x}_{hi}$$

Within this budget set of possible consumption bundles, the individual is presumed to choose his most preferred bundle. The most usual interpretation of "most preferred" in this context is that there is a preference ordering over all possible bundles, according to which, for every pair of bundles, one is preferred to the other or else the two are indifferent; and these pairwise judgments have the consistency property known to logicians as "transitivity;" thus, for example, if bundle A is preferred to bundle B and B to C, then A will be preferred to C. This "ordinalist" view of preferences was originally due to

Pareto and to Irving Fisher, about 1900, and represented an evolution from the earlier "cardinalist" position, according to which a measurable satisfaction or "utility" was associated with each bundle, and the consumer chose that bundle which maximized utility within the budget set. Obviously, a cardinal utility implies an ordinal preference but not *vice versa;* and if the only operational meaning of utility is in the explanation of consumer choice, then clearly two utility functions which defined the same preference ordering are operationally indistinguishable.[1]

The most preferred bundle then is a function, $x_{hi}(p_1, \ldots, p_n)$ of all prices. Notice that, from this viewpoint, all prices clearly enter into the determination of the demand for any one commodity. For one thing, the rise in any one price clearly diminishes the residual income available for all other commodities. More specifically, however, the demands for some commodities are closely interrelated with others; thus, the demand for gasoline is perhaps more influenced by the use of automobiles and therefore by their price than it is by its own price. The interrelation of all demands is clearly displayed here.

The characterization of consumer choice by optimization can, as we all know, be made more explicit. Let us recall Hicks's definition of the *marginal rate of substitution* between two commodities for any individual. For any given bundle, $(x_1^0, \ldots, x_n^0)$, consider all bundles indifferent to it, i.e., neither preferred to it nor inferior to it. If we hold all but two commodity quantities constant, say $x_k = x_k^0 (k \neq i, j)$, we can consider x_i as a function of x_j on this indifference surface. Then $-dx_i/dx_j$, evaluated at the point $x_i = x_i^0$, all i, is the marginal rate of substitution of commodity j for commodity i; it is, to a first approximation, the amount of commodity i that would be required to compensate for a loss of one unit of commodity j. The optimizing consumer will equate this marginal rate of substitution to the price ratio, p_j/p_i; for if the two were unequal, it would be possible to move along the indifference surface in some direction and reduce spending.

But since the marginal rate of substitution for any pair of commodities is equal to the price ratio for all individuals, it is also true that the marginal rate of substitution for any two commodities is the same for all individuals. This suggests in turn that there is no possibility that two or any number of individuals can gain by trading with each other after achieving a competitive equilibrium. The equality of the marginal rates of substitution means that a trade which would leave one individual on an indifference surface would do the same to the other. Hence, a competitive equilibrium satisfies the same kinds of conditions that are satisfied by a Pareto optimum.

(It will be observed that the stated conditions for a consumer optimum and for a Pareto optimum are first-order conditions in the differential calculus. Hotelling, Hicks, and Samuelson also developed the second-order conditions which distinguish maxima from minima and showed that these had important implications.)

Evaluation of the performance of an economy with regard to distributive justice was far less studied, not surprisingly, since the deepest philosophical issues are at stake. The Anglo-American tradition had incorporated in it one viewpoint, tacitly accepted though rarely given much prominence, the utilitarian views of Bentham and Sidgwick, given formal expression by Edgeworth. The criterion was the maximization of the sum of all individuals' utilities. This criterion only made sense if utility were regarded as cardinally measurable. With the rise of ordinalist doctrines, the epistemological basis for the sum-of-utilities criterion was eroded. It was to this issue that Bergson's famous paper (1938) was addressed. As already noted, a given preference ordering corresponds to many different utility functions. For any given set of preference orderings for the members of the economy, choose for each one of the utility functions which imply that preference ordering, and then the social welfare is expressed as some function, $W(U_1, \ldots, U_n)$ of the individual utilities. The function W will change appropriately if the utility indicator for the given preference orderings is changed, so that the entire representation is consistent with the ordinalist interpretation. However, the function W is not uniquely prescribed, as in the Edgeworth-Bentham sum of utilities, but is itself an expression of social welfare attitudes which may differ from individual to individual.

---

[1] The ordinalist view in fact only began to have wide currency in the 1930's, and indeed the treatments of Hicks and Samuelson, along with a paper of Hotelling's (1935), did much to make the ordinalist view standard. Interestingly enough, both Hicks and Samuelson have studied consumer choice by alternative axiom systems even weaker than ordinalism; see Hicks and Allen (1935) and Samuelson (1938).

So far, I have, for simplicity, spoken as if there were no production, an omission which must be repaired. A productive unit or firm is characterized by a relation between possible outputs and inputs. A firm may have, of course, more than one output. Then firm f may be characterized by its transformation surface, defined by an equation, $T(y_{f1}, \ldots, y_{fn}) = 0$, where y_{fi} is taken to be an output if positive and input if negative; the surface is taken to define the efficient possible input-output vectors for the firm, that is, those which yield maximum output of one commodity for given inputs and given outputs of other commodities. The optimizing behavior of the firm is taken to be the maximization of profit among the points on its transformation surface. Because of the sign conventions for inputs and outputs, the firm is seeking to maximize

$$\sum_{i=1}^{n} p_i y_{fi}$$

It is assumed in the treatment by Hicks and by Samuelson in the books referred to that the transformation surface is differentiable, so that the maximum-profit position is defined by suitable marginal equalities, and that the result is a function, $y_{fi}(p_1, \ldots, p_n)$ $(i=1, \ldots, n)$.

Two remarks should be made at this point: 1) Clearly, if all prices are multiplied by the same positive constant, the budget constraint for households is really unchanged, and hence so are the consumer demands. Similarly, the profits are multiplied by a positive constant, so that the profit-maximizing choice of a firm is unchanged. Hence, the functions $x_{hi}(p_1, \ldots, p_n)$ and $y_{fi}(p_1, \ldots, p_n)$ are homogeneous of degree zero in their arguments. 2) The firms' profits have to be treated as part of the income of the households that own them. This causes a modification of the previous budget constraint for the individual, which I shall not spell out in symbols here but will refer to below.

For any commodity i, there will be some demands and some supplies at any given set of prices. Following Hicks, we will speak of the *excess demand* for commodity i as the sum over all individuals and firms of demands and supplies, the latter being taken negative. The demand by individual h is $x_{hi}(p_1, \ldots, p_n)$, so that the total demand by all households is,

$$\sum_{h} x_{hi}(p_1, \ldots, p_n)$$

The supply by households is the aggregate amount they have to begin with, i.e.,

$$\sum_{h} \bar{x}_{hi}$$

Finally, the aggregate supply by firms is,

$$\sum_f y_{fi}(p_1, \ldots, p_n);$$

some firms may be demanders rather than suppliers, but the sign convention assures that the above sum gives the aggregate *net* supply by firms, i.e., after cancelling out demands by one firm which are supplied by another. Hence, the market excess demand for commodity i is,

$$z_i(p_1, \ldots, p_n) = \sum_h x_{hi}(p_1, \ldots, p_n) - \sum_h \bar{x}_{hi} - \sum_h y_{fi}(p_1, \ldots, p_n)$$

Since each term is homogeneous of degree zero, so is the total, z_i. Further, the satisfaction of the budget constraint for each individual also restricts the excess demand functions. Since for each individual the monetary value of expenditures planned at any set of prices equals the monetary value of his initial endowments plus his share of the profits, we have in the aggregate that the money value of planned expenditures by all households equals the money value of total endowments plus total profits, or,

$$\sum_h \sum_{i=1}^n p_i x_{hi}(p_1, \ldots, p_n) =$$

$$\sum_h \sum_{i=1}^n p_i \bar{x}_{hi} + \sum_f \sum_{i=1}^n p_i y_{fi}(p_1, \ldots, p_n)$$

or, from the definition of excess demand,

$$\sum_i p_i z_i(p_1, \ldots, p_n) \equiv 0$$

where the identity symbol reminds us that this relation, called by Lange (1942) Walras' Law, holds for all values of the prices.

The general equilibrium of the economy is then the set of prices which equate all excess demands to zero,

$$z_i(p_1, \ldots, p_n) = 0 (i = 1, \ldots, n)$$

These appear to be n equations in n unknowns; but there are two off-setting complications in the counting. On the one hand, since the equations are homogeneous, no solution can be unique, since any positive multiple of all prices is also a solution. In effect, the equations really only determine the $n-1$ price ratios. On the other hand, the equations are not independent; if $n-1$ are satisfied, then the nth must be by Walras' Law.

III. The Need for Further Development

There were, however, several directions in which the structure of general equilibrium theory was either incomplete or inconsistent with doctrines which had strong currency in economic theory.

(1) There was no proof offered that the system of equations defining general equilibrium had a solution at all; that is, it was not known that there existed a set of prices which would make excess demand zero on every market. This was the most serious unresolved problem.

(2) The assumptions on production were not the same as those used in the analysis of production itself. In the latter, a common, though not universal, assumption was that of constant returns to scale; if any production process can be carried out, with given inputs and outputs, then the process can be carried out at any scale. That is, if the inputs are all multiplied by the same positive number, then it is possible to produce the same multiple of all the outputs. But in this case, there cannot be a unique profit-maximizing position for any set of prices. For suppose there were a position which yielded positive profits. Then doubling all inputs and outputs is feasible and yields twice as great profits. Hence, there would be no profit-maximizing position, since any one could be improved upon. On the other hand, zero profits can always be obtained by having no inputs and no outputs. It can be concluded that, if prices are such that there is some profit-maximizing set

of inputs and outputs not all zero, the corresponding profits must be zero, and the same profits can be achieved by multiplying all inputs and outputs by any positive number.

Therefore, under constant returns to scale, there is never a single-valued function, $y_{fi}(p_1, \ldots, p_n)$ defining inputs and outputs as a function of prices; rather, for any given set of prices, either there is no profit-maximizing input-output vector or else there is a whole ray of them. But then the notion of equating supply and demand must be redefined.

Of somewhat lesser importance in this regard is the fact that the transformation surface need not be differentiable in very plausible circumstances. A frequently held view was that production of a given output required prescribed amounts of each input; in some circumstances, at least, it is impossible to reduce the need for one input by increasing the amount of another. This is the fixed-coefficient technology. In this case, it can easily be seen that though the transformation surface is well defined, it is not differentiable but has kinks in it.

(3) The relation between Pareto-efficient allocations and competitive equilibria was less clearly formulated than might be desired. What had really been shown was that the necessary first-order conditions for Pareto efficiency were the same as the first-order conditions for maximization by firms and individuals when the entire economy is in a competitive equilibrium.

(4) Actually, the condition for individual optimization (equating of marginal rates of substitution to price ratios) required some modification to take care of corner maxima. It is obvious to everyday observation that for each individual there are some (indeed, many) commodities of which he consumes nothing. Similarly, for every firm, there are some commodities which are neither inputs to nor outputs of it. But then the argument that the marginal rate of substitution must equal the price ratio for each individual breaks down. For consider an individual for whom the marginal rate of substitution of commodity j for commodity i is *less* than the price ratio, p_j/p_i, but the individual consumes nothing of commodity j. A small increase in the consumption of j with a compensating decrease in i to stay on the same indifference surface would involve an increase in costs. The only way to achieve a decrease in cost without moving to a less preferred position would be to decrease the consumption of j; but this is impossible, since consumption cannot fall below zero. It is true, however, that the marginal rate of substitution of j for i cannot *exceed* the price ratio.

Similarly, if one individual consumes nothing of commodity j, it is possible to have Pareto efficiency with his marginal rate of substitution of j for i less than that for some other indiviual. Since marginal rates of substitution do not have to be equated across individuals either for competitive equilibria or for Pareto-efficient allocations, the relation between the two concepts was seen to need further study.

(5) Still another question is whether supply and demand are necessarily equal. Clearly, demand cannot exceed supply, for there would have to be unfulfilled demands. But as we look around us, we see that there are goods, i.e., flows which we prefer to have, which nevertheless are so abundant that we have no desire for more. Air and sunlight come immediately to mind. Characteristically, such highly abundant goods are free; no price is charged for their use.

This elementary observation has been made a number of times by economists. A distinction was drawn between *scarce goods* and *free goods*, the former alone being the proper subject matter of economics. But it is easy to see from a mathematical viewpoint that the classification of goods in this way is not a given but depends on those parameters of the system which govern tastes, technology, and initial supplies. Suppose, for example, that we have two commodities, A and B, which serve as factors of production only. Suppose further it so happens that the two factors are always used together and always in the same proportion, say, one unit of A with two units of B. Finally, suppose that A and B are not themselves produced goods but are natural resources available in equal quantities. Then clearly commodity B is the bottleneck; commodity A is a free good in the usual economic sense, since a small change in the quantity available would have no effect on production. But this classification of the two goods into free and scarce is relative to the technology and to the initial supplies of the two goods. If a technological innovation reduced the need for B so that one unit of A required the cooperation of less than one unit of B, B would become the free good, and A, the scarce one; and the same would happen if the initial supply of A were reduced, perhaps by some catastrophe, to less than half of that of B.

The conditions for equilibrium then have to be modified. We require now that excess demand be nonpositive and that, for any commodity for which it is negative, the price be zero. In symbols,

$$z_i(p_1, \ldots, p_n) \leqq 0 (i = 1, \ldots, n),$$

if

$$z_i(p_1, \ldots, p_n) < 0, \quad \text{then } p_i = 0$$

The commodities for which the inequality holds are the free goods. Equilibria in which there are free goods are referred to as *corner equilibria.*

The problem just raised illustrates a general tendency in the evolution of general equilibrium theory for a shift from a local to a global analysis. If we consider small shifts in the parameters which determine tastes, technology, and initial supplies, the classification of goods into free and scarce remains unchanged. Hence, from a local viewpoint, the list of scarce goods could legitimately be taken as given. We need not debate here the relative virtues of local and global analysis: clearly a global analysis is always preferable if it is possible, but a local analysis will normally produce more specific implications. But it turns out that the first of the problems raised, that of the existence of equilibrium prices, cannot be handled at all except from a global viewpoint; and the realization of the possibility of corner equilibria turned out to be an indispensable step in the development of an existence proof.

To avoid a misinterpretation of this list of the needs for further development, two points should be stressed: 1) the general aims and structure of general equilibrium theory have remained those already set forth by Hicks, and the subsequent development would have been impossible and indeed meaningless except on his foundations; 2) I have summarized here only the most general and foundational aspects of the work of Hicks and Samuelson, since those are most relevant for my present purpose, but the primary interest of both was rather in the laws of working of the general equilibrium system, results not summarized above, than in the questions of existence and the like.

IV. The German Language Literature

We must turn from the Anglo-American work to a variant strand of neoclassical thought, published primarily in German, and written to a considerable extent by mathematicians rather than economists. The whole literature might be described as an extended commentary on a formulation of general equilibrium theory by Cassel (1918), a statement rather different in nature from that of Hicks. In particular, maximizing behavior hardly appeared in Cassel's model. With regard to individual consumers, Cassel also assumed that the demand of individual households was a function of prices; he did not, however, seek to derive this demand from a preference or utility maximization. With regard to production, he assumed a fixed-coefficient technology, so that there was in effect no scope for profit maximization by firms; the demands for inputs were completely defined by the outputs, independent of prices. More explicitly, Cassel differentiated commodities into produced goods and primary factors, the two classes being assumed distinct. Individuals owned initially only primary factors, and they demand only produced goods. Produced goods were made by inputs of primary factors; let a_{ij} be the amount of factor j used in the production of one unit of good i. Let P be the set of produced goods, F, that of factors.

At any set of prices, the total demand for produced good i is,

$$\sum_h x_{hi}(p_1, \ldots, p_n) = x_i;$$

the demand for factor j by the industry producing good i is then $a_{ij}x_i$, and the total demand for factor j is obtained by summing this demand over all producing industries. On the other hand, the initial supply of factor j is $\sum_h \bar{x}_{hj}$, so that the condition for equality of supply and demand for factor j is,

$$\sum_h \bar{x}_{hj} = \sum_{i \in P} a_{ij}x_i$$

As j varies over F, we have a system of linear equations in the x_i's. Now von Stackelberg (1933) observed that this system might easily have no solution, for example, if there are more factors than produced goods.

Cassel completed the system by using the condition that, under constant returns to scale, there must be zero profits. Then, for each produced good, the price must equal the cost of the factors used in making one unit, or

$$p_i = \sum_{j \in F} a_{ij}p_j (i \in P)$$

About contemporaneously with von Stackelberg, Neisser (1932) showed that it could easily be true that the complete Cassel system could be satisfied only if some factor prices were negative.

It was at this point that the Viennese banker and amateur economist, K. Schlesinger (1933–34), decisively affected the subsequent discussion. He observed that the criticisms raised by von Stackelberg and by Neisser could be met by recognizing the possibility of corner equilibria, particularly with regard to primary factors. Some may simply be superfluous and have to be regarded as free goods. Thus, the equality of supply and demand for factors has to be replaced by the following conditions:

$$\sum_h \bar{x}_{hj} \geqq \sum_{i \in P} a_{ij}x_i, \quad p_j = 0$$

if the strict inequality holds ($j \in F$). With this amendment, Schlesinger conjectured, it could be shown that there existed an equilibrium in which all prices are nonnegative. He interested the mathematician Abraham Wald in this problem, and the latter showed in a brilliant series of papers (1933–34, 1934–35), summarized in 1936, that equilibrium indeed existed, though rather strong assumptions had to be made and the analysis was confined to variations of the Cassel model. Wald's reasoning was formidably complex, and his work published in a German language mathematics journal; it was only some ten years later that American mathematical economists began to be aware of it.

Within the same period, the mathematician John von Neumann published a paper (1937) which had in the longer run a deeper impact, though its subject matter was less relevant. This was a development of Cassel's model of steady growth of the economy. The aim was to show the existence of a growth path with maximum proportional expansion in all commodities. From an economic point of view the model was somewhat strange in that there was no consumption at all; the outputs of one period were inputs into activities which generated the outputs of the next period. There were three noteworthy points which had great influence on the development of general equilibrium theory: 1) The structure of production was characterized in a novel way. It was assumed that there were a fixed set of activities, each being characterized by a vector of possible inputs and outputs and each being technologically capable of operation at any scale. This generalized the fixed-coefficient model, in which there was one activity for each output. The feasible combinations of activities were those for which the total usage of each input did not exceed the amount available from previous production. 2) The maximum growth path could be characterized as a sort of competitive equilibrium, in the sense that it was mathematically possible and meaningful to introduce a new set of variables, which could be regarded as prices. Any activity that was run at all yielded zero profits; other activities yielded zero or negative profits. Hence, the choice of activity levels could be described as profit-maximizing, where the maxima may involve some corners. Further, the price of any commodity for which the demand as input fell short of the amount available had to be zero; hence, the competitive equilibrium could require corners. 3) The method of proof of the existence of prices and relative quantities which yielded a maximum growth rate required the use of a tool from combinatorial topology, a generalization of Brouwer's fixed point theorem. From a mathematical viewpoint, the existence of equilibrium in the von Neumann growth model was a generalization of the minimax theorem for zero-sum two-person games, which von Neumann had studied a few years earlier. The interest in game theory following the publication of the great book of von Neumann and O. Morgenstern (1944) was a strong collateral force in introducing new mathematical techniques, particularly in the theory of convex sets, into general equilibrium theory.

A simplification of von Neumann's fixed point theorem was developed a few years later by S. Kakutani (1941) and has become the standard tool for proving existence theorems. Let us review briefly the fixed point theorems of Brouwer and Kakutani. Recall that a set of points is said to be compact if it is closed and bounded and to be convex if every line segment joining two points of the set lies entirely within the set. Let **C** be a compact convex set. Let $f(x)$ be a vector function which assigns to every point of **C** a point of **C**. Then Brouwer's theorem asserts that if the mapping $f(x)$ is continuous, then there is at least one point, x^*, which is mapped into itself, i.e., for which $f(x^*) = x^*$.

In the indicial notation which we have used hitherto, we have n real-valued functions $f_i(x_1, \ldots, x_n)$ of n variables. If these functions are continuous and if the point $(f_1, \ldots, f_n)$ lies in some compact convex set **C** whenever $(x_1, \ldots, x_n)$ lies in that set, then the system of equations, $f_i(x_1, \ldots, x_n) = x_i$, has at least one solution in **C**.

The relevance of such a mathematical tool to the problem of existence is obvious. However, we have already noted above that once we permit constant returns to scale, we have to allow for the possibility that the profit-maximizing choice of production process may be a whole set, all equally profitable, for some given set of prices. Hence, instead of dealing with functions, we need to concern ourselves with the more general notion of a point-to-set mapping, or *correspondence*, as it is sometimes termed. Kakutani's theorem deals with this more general situation. To every point $x=(x_1, \ldots, x_n)$ in a compact convex set C, we associate a subset of C, say $\Phi(x)$. We say that x^* is a *fixed point* of this correspondence if the point x^* belongs to the set associated with x^*, i.e., to $\Phi(x^*)$.

Kakutani's theorem tells us that such a fixed point will exist if two conditions are fulfilled: for each x, $\Phi(x)$ is a convex set; and as x varies, $\Phi(x)$ is continuous in a certain sense, more technically, that it has the property known as *upper semi-continuity*.

V. Pareto Efficiency, Competitive Equilibrium, and Convexity

My own interest first centered on the relations between Pareto efficiency and competitive equilibrium. In particular, there was considerable discussion among economists in the late 1940's about the inefficiencies resulting from rent control and different proposals for arriving at the efficiency benefits of a free market by one or another transition route. Part of the informal efficiency arguments hinged on the idea that under rent control people were buying the wrong kind of housing, say, excessively large apartments. It struck me that an individual bought only one kind of housing, not several. The individual optima were at corners, and therefore one could not equate marginal rates of substitution by going over to a free market. Yet diagrammatic analysis of simple cases suggested to me that the traditional identification of competitive equilibrium and Pareto efficiency was correct but could not be proved by the local techniques of the differential calculus.

I soon realized that the theory of convex sets, and, in particular, the separation theorem, was the appropriate tool. Start with a Pareto-efficient allocation, and consider all logically possible allocations which would be preferred to it by everyone. Of course, no such allocation can be feasible; otherwise the allocation we started with would not be Pareto efficient. Each such allocation is a statement of demand or supply of each commodity by each individual or firm. Hence, by adding up over individuals and firms, with appropriate attention to signs, we can define the excess demand for each commodity. Let Z be the set of all excess demand vectors $(z_1, \ldots, z_n)$ generated this way. Since they are all infeasible, it must be true for each one that there is positive excess demand for at least one commodity. In the language of set theory, the set Z is disjoint from the nonpositive orthant, i.e., the set of vectors $(z_1, \ldots, z_n)$ such that $z_i \leqq 0$ for all i.

The separation theorem for convex sets asserts that if two convex sets are disjoint, there is a hyperplane which separates them, so that one set is on one side and the other set on the other. In symbols, if C_1 and C_2 are disjoint convex sets in n-dimensional space, there exists numbers p_i $(i=1, \ldots, n)$, not all zero, c, such that $\sum_{i=1}^{n} p_i x_i \geqq c$ for all $x=(x_1, \ldots, x_n)$ in C_1, $\sum_{i=1}^{n} p_i x_i \leqq c$ for all x in C_2. Let us apply this theorem to the present case. The nonpositive orthant is obviously a convex set; let us assume for the moment that Z is convex. Then we can find numbers p_i $(i=1, \ldots, n)$, not all zero, c such that,

$$\sum_{i=1}^{n} p_i z_i \geqq c \qquad \text{for } z=(z_1, \ldots, z_n) \text{ in } Z,$$

$$\sum_{i=1}^{n} p_i z_i \leqq c \qquad \text{if } z_i \leqq 0 \text{ for all } i$$

From the second condition, it can easily be seen that we cannot have $p_i < 0$ for any i. Hence, p_i is nonnegative for all i and (since there is at least one nonzero p_i) positive for at least one i. This is customarily expressed by saying that the vector $p = (p_1, \ldots, p_n)$ is *semi-positive*.

It follows that,

$$\sum_{i=1}^{n} p_i z_i \leqq 0 \qquad \text{if } z_i \leqq 0 \text{ for all } i,$$

and therefore we can assume without loss that $c \leqq 0$. On the other hand, if we set $z_i = 0$ for all i, we see that $c \geqq 0$. Hence, we can set $c = 0$.

The conditions for a Pareto-efficient allocation then become,

$$\sum_{i=1}^{n} p_i z_i \geqq 0 \qquad \text{for } z \text{ in } \boldsymbol{Z},\ p \text{ semi-positive}$$

Let $z^0 = (z_1^0, \ldots, z_n^0)$ be the vector of excess demands defined by the Pareto-efficient allocation under consideration. It is feasible, so that $z_i^0 \leqq 0$, all i, and therefore,

$$\sum_{i=1}^{n} p_i \overset{0}{z_i} \leqq 0$$

Now assume, as is usually reasonable, that there are points in $\boldsymbol{Z}$ as close as one wishes to z^0. Then clearly we must have,

$$\sum_{i=1}^{n} p_i \overset{0}{z_i} = 0,$$

and hence, since $z_i^0 \leqq 0$, all i, $p_i \geqq 0$, all i, that,

$$\text{if } \overset{0}{z_i} < 0, \qquad p_i = 0$$

We begin to see that a Pareto-efficient allocation is an equilibrium of supply and demand in the generalized sense which includes corners.

We also see that,

$$\sum_{i=1}^{n} p_i (z_i - \overset{0}{z_i}) \geqq 0 \qquad \text{for } z \text{ in } \boldsymbol{Z}$$

Let us go back to the definition of excess demand, as a sum of individual and firm demands and supplies.

$$z_i = \sum_h x_{hi} - \sum_h \bar{x}_{hi} - \sum_f y_{fi}$$

where $y_f(y_{f1}, \ldots, y_{fn})$ is a technologically possible vector of inputs and outputs for firm f and $x_h = (x_{h1}, \ldots, x_{hn})$ is a possible vector of consumptions for individual h. In particular, the excess demands defined by the Pareto-efficient allocation can be written in this form,

$$\overset{0}{z_i} = \sum_h \overset{0}{x_{hi}} - \sum_h \bar{x}_{hi} - \sum_f \overset{0}{y_{fi}}$$

and then, if z belongs to $\boldsymbol{Z}$, we must have, for each h, that the consumption vector of individual h, $(x_{h1}, \ldots, x_{hn})$ is preferred to that under the Pareto-efficient allocation $(x_{h1}^0, \ldots, x_{hn}^0)$. Then,

$$\sum_h \left(\sum_{i=1}^{n} p_i x_{hi} - \sum_{i=1}^{n} p_i \overset{0}{x_{hi}} \right) - \sum_f \left(\sum_{i=1}^{n} p_i y_{fi} - \sum_{i=1}^{n} p_i \overset{0}{y_{fi}} \right) \geqq 0$$

if, for each h, x_h is preferred by individual h to x_h^0.

Now the elementary point about this inequality is that the variable vectors x_h, y_f are independent of each other. It is not hard to see that this inequality can hold only if it holds for each individual and each firm separately. For a firm f, this means that,

$$\sum_{i=1}^{n} p_i \overset{0}{y_{fi}} \geqq \sum_{i=1}^{n} p_i y_{fi} \qquad \text{for all possible } y_f,$$

that is, if we interpret the p_i's as prices, each firm is maximizing its profits. The corresponding interpretation for individuals is somewhat less simple; it is that the consumption vector prescribed by the given Pareto-efficient allocation is the cheapest way of deriving that much satisfaction.

Taken altogether, it has been shown that if Z is a convex set, the Pareto-efficient allocation can be achieved as a competitive equilibrium of the market, in the sense that prices and a suitable initial allocation of resources can be found such that each individual is achieving his satisfaction level at minimum cost, each firm is maximizing profits, and the markets are all in equilibrium in the generalized sense which permits corner equilibria.

The need to assume that Z is convex puts in sharper focus the convexity assumptions which had always implicitly underlain neoclassical theory. The convexity of Z could be derived from the following two assumptions: 1) for each individual, the set of consumption vectors preferred to a given vector is convex; 2) for each firm, the set of technologically possible vectors is convex.

The result states that, under suitable convexity conditions, a necessary condition for an allocation to be Pareto efficient is that it be realizable in the market as a competitive equilibrium. A by-product of the investigation was the proof of the converse theorem: a competitive equilibrium is always Pareto efficient, and this theorem is true without any convexity assumption.

These results were embodied in Arrow (1951a). But the idea that the theory of convex sets was the appropriate tool was clearly in the air. While I was working at Stanford, Gerard Debreu (1951) obtained very much the same results at the Cowles Commission for Research in Economics at Chicago.

VI. The Existence of Competitive Equilibrium

Again working independently and in ignorance of each other's activities, Debreu and I both started applying Kakutani's fixed point theorem to the problem of existence. In this case, we exchanged manuscripts in sufficient time to realize our common efforts and also to realize the need for relaxing an excessively severe assumption we had both made (Arrow and Debreu (1954)).

An essential precondition for our studies was the basic work of Tjalling Koopmans (1951) on the analysis of production in terms of activity analysis. In this he extended von Neumann's work into a systematic account of the production structure of the economy. He saw it as a set of activities, each of which could be operated at any level but with the overall levels constrained by initial resource limitations. The crucial novelty was the explicit statement of the assumptions which insured that the feasible set of outputs would be bounded for any finite set of initial resources. It turned that this limitation is a "global" property. That is, conditions on the nature of individual activities (for example, that every activity had to have at least one input) were not sufficient to insure the boundedness of the economy as a whole. It was necessary to require that no combination of activities as a whole permitted production without inputs.

The first question is the definition of equilibrium when the behavior of firms is described by a correspondence rather than a function. For simplicity, I will continue to assume that the decisions of individual consumer h can be represented by single-valued functions of prices, $x_h(p)$. A set of prices defines a competitive equilibrium if supply and demand balance on each market, including the possibility of corners, with *some* choice of the profit-maximizing input-output vector for each firm. Formally, we will say that a price vector p^*, an input-output vector y_f^* for each firm, and a consumption vector, $x_h^* = x_h(p^*)$, for each individual together constitute a competitive equilibrium if the following conditions hold:

(a) p^* is semi-positive;

(b) for each commodity i,

$\sum_h \bar{x}_{hi} + \sum_f y^*_{fi} \geqq \sum_h x^*_{hi}$;

(c) for any commodity for which the strict inequality holds in (b), we must have $p^*_i = 0$;

(d) y^*_f is one of the input-output vectors which maximizes profits, $\sum_{i=1}^n p^*_i y_{fi}$, among all the input-output vectors technologically possible for firm f.

It is, of course, understood that the demand function for individuals, $x_h(p)$ is defined, as before, as the most preferred consumption pattern consistent with the budget constraint,

$$\sum_{i=1}^n p_i x_{hi} \leqq \sum_{i=1}^n p_i \bar{x}_{hi}:$$

for the present purposes, I will ignore the possibility that individuals' incomes also include profits; this modification can be handled at the cost of some analytic complexity but no true difficulty.

It will be assumed 1) that the set of possible input-output vectors for any firm is convex, and 2) the individual demand functions are continuous; this assumption will be discussed again below.

Since the total production possibilities of the economy are bounded, it can be shown there is no loss of generality in assuming the set of possible input-output vectors for each firm is bounded (actually, we assume the set to be compact). Then for any set of prices there is at least one profit-maximizing input-output vector, but in general there may be a whole set of them, say $\boldsymbol{Y}_f(p)$. However, this set is certainly convex and further, as p varies, the correspondence so defined is upper semi-continuous.

Define an excess demand correspondence as follows: For each f, consider any possible selection of a vector y_f from the profit-maximizing correspondence $Y_f(p)$. For each such selection for each firm, form the excess demand for each commodity for the entire economy,

$$z_i = \sum_h x_{hi}(p) - \sum_f y_{fi} - \sum_h \bar{x}_{hi}$$

Let $\boldsymbol{Z}(p)$ be the set of all vectors $(z_1, \ldots, z_n)$ which can be formed by all possible selections of the vectors y_f from the profit-maximizing correspondence $Y_f(p)$, the selections for different firms being made independently of each other. It is not hard to show that $\boldsymbol{Z}(p)$ is convex for each p and is an upper semi-continuous correspondence for p as a variable. It is also true and important that Walras' Law holds; that is, if z belongs to $\boldsymbol{Z}(p)$, then

$$\sum_{i=1}^n p_i z_i = 0$$

The correspondence $\boldsymbol{Z}(p)$ assigns to each price vector a set of excess demands; an equilibrium price vector p^* would be one such that $\boldsymbol{Z}(p^*)$ has at least one element for which $z_i \leqq 0$, all i. We now introduce a mapping from excess demands; very roughly, we want low excess demands to have low or more precisely zero prices. Since the whole system is homogeneous of degree zero in the prices, the general level of the prices can be set arbitrarily with no loss of generality. It will be assumed then that

$$\sum_{i=1}^n p_i = 1$$

Since prices are semi-positive, it is also assumed that $p_i \geqq 0$, all i; the set of price vectors satisfying these conditions will be denoted by $\boldsymbol{P}$, the price simplex. Then we define the following correspondence, assigning to each vector of excess demands, a subset of the price simplex: for any $z = (z_1, \ldots, z_n)$, let $\bar{z}$ be the largest of the components z_i; then define $P(z)$ to be the set of price vectors in the unit simplex for which $p_i = 0$ for all commodities i for which $z_i < \bar{z}$. In words, total prices must add up to one, but this total is to be distributed only over those commodities with maximum excess demand. This rule is somewhat artificial, but it suffices for the proof.

Consider the set of all pairs (z, p) of vectors, one an excess demand vector and one a price vector. To any such pair we assign a set of pairs, $\boldsymbol{Z}(p) \times \boldsymbol{P}(z)$ (for any pair of sets, S, T, the notation $S \times T$ means the set of ordered pairs of vectors obtained by taking any vector from S followed by any vector from T). With some further argument, it can be shown that Kakutani's theorem applies. The mapping of pairs has a fixed point, (z^*, p^*) belonging to $\boldsymbol{Z}(p^*) \times \boldsymbol{P}(z^*)$. By definition,

$$z^* \in \boldsymbol{Z}(p^*), \qquad p^* \in \boldsymbol{P}(z^*)$$

Let $\bar{z}^*$ be the largest component of $z^* = (z_1^*, \ldots, z_n^*)$. Then $p_i^* = 0$ for $z^* < \bar{z}^*$. Therefore,

$$\overset{*}{p}_i \overset{*}{z}_i = \overset{*}{p}_i \bar{z}^*$$

By Walras' Law,

$$0 = \sum_{i=1}^{n} \overset{*}{p}_i \overset{*}{z}_i = \sum_{i=1}^{n} \overset{*}{p}_i \bar{z}^* = \bar{z}^* \sum_{i=1}^{n} \overset{*}{p}_i = \bar{z}^*$$

since the largest excess demand is zero, all excess demands are nonpositive, and therefore p^* is indeed an equilibrium price vector.

Many variations of this argument are possible and are illuminating in different ways. Independently of my work with Debreu, Lionel McKenzie (1954) proved the existence of equilibrium; he simply assumed the existence of supply and demand functions rather than analyzing them in terms of the underlying production and consumption structures. For systematic presentations of the existence theorems for competitive equilibrium, see Debreu (1959) and Arrow and Hahn (1971, chs. 2–5).

There is one loose end that should now be picked up. It has been assumed that the demand functions of the individual are continuous. But one of the surprising discoveries that Debreu and I made in the course of our study was that even under all the usual strong assumptions about the behavior of individuals, this cannot be true everywhere in the price simplex except under very artificial conditions. The trouble is that the individual's income also depends upon prices, and if the prices of those commodities which the individual owns originally fall to zero, his income falls to zero. When some prices and income are zero, however, the demand for the now-free goods may jump discontinuously. To illustrate, suppose an individual owned initially only one good, say, labor. So long as the price of that good was positive, he might retain some for his own use, but in any case could never consume more than he had initially. But when the price fell to zero, he could demand the same labor from others and in any amount he chooses.

The existence of competitive equilibrium then does depend on assumptions which insure that for each individual there is at least one commodity he owns initially which is bound to have positive value. I will not state these assumptions here; the original set in Arrow-Debreu has been refined through the work of Gale (1957), McKenzie (1959, 1961), and Arrow and Hahn (1971, ch. 5, section 4).

VII. General Equilibrium and Uncertainty

Once the broad approach to the analysis of existence was set, it could be applied in many different directions. One was the analysis of models which represented in one way or another imperfections in the competitive system. The requirement of proving an existence theorem in each case leads to the need for a rigorous spelling out of assumptions, a requirement which seems to be proving very fruitful. Much of this work is now going on, in such areas as the analysis of futures markets, expectations, and monetary theory, but time does not permit comment on what is in any case a rapidly changing field.

Another approach is to retain the competitive assumptions but interpret them in new contexts. One example of this is the extension of general equilibrium theory to uncertain outcomes (Arrow (1953); Debreu (1959, ch. 7)). Suppose there is some uncertainty in production due, for example, to the weather. One type of weather will benefit one kind of producer and injure another, while another type will do the opposite. If we assume that individuals are averse to risk, there is room for a mutually profitable trade in insurance. Even apart from risk aversion, individuals and firms in planning for an uncertain future may want to make sure that their demands and outputs are mutually compatible. Thus, if there is uncertainty about the supply of grain, a miller may prefer to make future contracts for labor contingent on that uncertainty.

We take from the theory of probability the concept of a *state of the world*, which is a description of the world so precise that it completely defines all initial holdings of goods and all technological possibilities. Uncertainty is not knowing which state will in fact hold. The initial holdings of commodity i by individual h if state s should hold can be designated by $\bar{x}_{his}$. Similarly, the set of possible input-output vectors for a firm may depend upon the state s; let $y_{fs}=(y_{f1s}, \ldots, y_{fns})$ be a possible input-output vector for firm f if s is the state.

The feasibility of any allocation will then depend upon the state s, and therefore commitments to consumption and production must vary similarly. Hence the decision by any individual must be a separate vector $x_{hs}=(x_{h1s}, \ldots, x_{hns})$ for each state s. But clearly it is optimal for all concerned to make all these decisions simultaneously, in advance of knowing which state of the world will in fact prevail; it is this advance decision which permits the possible gains from insurance, from the reduction in risk bearing. Hence, we should really think of the vector x_h, which, for fixed h, contains components x_{his} where i and s range over commodities and states of the world, respectively.

What we are led to is considering the same physical commodity in different states of the world as economically different commodities. The procedure is exactly analogous to Hick's analysis of present and future goods (1939); the same physical commodity at different points of time define different commodities.

The whole previous analysis can then be applied, with a suitable reinterpretation. Commodities in the ordinary sense are replaced by *contingent commodities*, promises to buy or sell a given commodity if, and only if, a certain state of the world occurs. The market will then determine contingent prices. Clearing of the markets means clearing of the contingent markets; the commitments made are sufficiently flexible so that they can always be satisfied.

It should be noted that preference orderings over vectors of contingent commodities contain elements of judgment about the likelihoods of different states of the world as well as elements of taste in the ordinary sense. Other things being equal, one will invest less heavily in a demand contingent upon a state deemed unlikely.

One can work out the implications of this model. Clearly, the contingent commodities called for do not exist to the extent required, but the variety of securities available on modern markets serves as a partial substitute. In my own thinking, the model of general equilibrium under uncertainty is as much a normative ideal as an empirical description. It is the way the actual world differs from the criteria of the model which suggests social policy to improve the efficiency with which risk bearing is allocated.

In fact, it is not a mere empirical accident that not all the contingent markets needed for efficiency exist, but a necessary fact with deep implications for the workings and structure of economic institutions. Roughly speaking, information about particular events, even after they have occurred, is not spread evenly throughout the population. Two people cannot enter into a contract contingent on the occurrence of a certain event or state if only one of them in fact will know that the event has occurred. A particular example of this is sometimes known as "moral hazard" in the insurance and economic literature. The very existence of insurance will change individual behavior in the direction of less care in avoiding risks. The insurance policy that would be called for by an optimal allocation of risk bearing would only cover unavoidable risks and would distinguish their effects from those due to behavior of the individual. But in fact all the insurer can observe is a result, for example, a fire or the success or failure of a business, and cannot decompose it into exogenous and endogenous components. Contingent contracts, to speak generally, can be written only on mutually observed events, not on aspects of the state of the world which may be known to one but not both of the parties.

Although I cannot argue the point here, I would hold that the allocational difficulties arising from the inequality in information are of importance in such diverse fields as medical care and racial discrimination (see Arrow (1963a, 1972)). The difficulty of achieving optimal allocation of risk bearing because of differences in information was first stated in a general form by Roy Radner (1968).

VIII. The Theory of Social Choice

General competitive equilibrium above all teaches the extent to which a social allocation of resources can be achieved by independent private decisions coordinated through the market. We are assured indeed that not only can an allocation be achieved, but the result will be Pareto efficient. But, as has been stressed, there is nothing in the process which guarantees that the distribution be just. Indeed, the theory teaches us that the final allocation will depend on the distribution of initial supplies and of ownership of firms. If we want to rely on the virtues of the market but also to achieve a more just distribution, the theory suggests the strategy of changing the initial distribution rather than interfering with the allocation process at some later stage.

Thus even under the assumptions most favorable to decentralization of decision making, there is an irreducible need for a social or collective choice on distribution. In point of fact, there are a great many other situations in which the replacement of market by collective decision making is necessary or at least desirable. In their different ways, both political scientists and economists have discussed the necessary role of the state. Among economists, these discussions have revolved around the concepts of externalities, increasing returns, and market failure; the clarification and application of these ideas have been among the major achievements of modern economic thought, but I have time now merely to recall them to you as helping to create the need for normative and descriptive analysis of collective decision making.

In the context of social choice, each individual may be assumed to have a preference ordering over all possible social states. This ordering expresses not only his desire for his own consumption but also social attitudes, his views on justice in distribution or on benefits to others from collective decisions. The ordinalist viewpoint forbids us from ascribing a definite quantitative expression to this preference, at least a quantitative expression which would have any interpersonal validity.

Classical utilitarianism specifies that choices among alternative social states be judged in terms of their consequences for the members of the society; in the present terminology, this means in terms of the individual preference scales for social choices. This is obviously not a sufficient basis for choice in view of the diversity of individual preferences. It is implicit in classical utilitarianism and explicit in Bergson's work that there is a second level at the individual judgments are aggregated into what might be termed a welfare judgment.

Thus the formation of welfare judgments is logically equivalent to what I will call a *constitution*. Specifically, a constitution is a rule which associates to each possible set of individual preference orderings a social choice rule. A social choice rule, in turn, is a rule for selecting a socially preferred action out of any set of alternatives which may be feasible.

So far, I would hold that the description of a constitution is a tautology, at least if we start from the view that social choice has to be based on the individual preference orderings. The real question is what conditions are to be imposed on the constitution.

One condition, which is already contained in Bergson's work, is that for any given set of individual preferences, the social choice rule defined by them shall satisfy the technical conditions of an ordering; that is, that all possible alternative social states should be capable of being ranked and then the social choice from any particular set of alternatives should be the most preferred alternative, according to the ordering, in the available set. This is sometimes called the condition of Collective Rationality.

A second condition, again in agreement with Bergson, is the Pareto principle; the social choice process shall never yield an outcome if there is another feasible alternative which everyone prefers according to his preference ordering.

A third hardly controversial condition is that of Non-Dictatorship; the constitution shall not be such that there is an individual whose preferences automatically become those of society regardless of anyone else's preferences.

The fourth condition which I have suggested, that of the Independence of Irrelevant Alternatives, is more disputable, though I would argue that it has strong pragmatic justification: the social choice made from any set of alternatives will depend on only the orderings of individuals among alternatives in that set. To see what is at stake, suppose that a society has to make a choice among some alternatives and does so. After the decision is made, an alternative which has not previously been thought of is mentioned as a logical possibility, although it is still not feasible. The individuals can expand their preference orderings to place this new alternative in its place on their ranking; but should this preference information about an alternative which could not be chosen in any case affect the previous decision?

Any form of voting certainly satisfies the condition of Independence of Irrelevant Alternatives; the preferences of voters as between candidates and noncandidates or as between noncandidates are, of course, never asked for or taken into account.

It turns out (Arrow (1951b, 1963b)) that these four reasonable sounding requirements are contradictory. That is, if we devise any constitution, then it is always possible to find a set of individual orderings which will cause the constitution to violate one of these conditions. In one special form, this paradox is old. The method of majority voting is an appealing method of social choice. Like any other voting method, it satisfies Independence of Irrelevant Alternatives and certainly the Pareto principle and the condition of Non-Dictatorship. But as Condorcet pointed out as far back as 1785, majority voting may not lead to an ordering. More specifically, intransitivity is possible. Consider the following example. There are three alternatives x, y, and z, among which choice is to be made. One-third of the voters have the ranking x, y, z; one-third, the ranking y, z, x; and one-third, the ranking z, x, y. Then a majority of the voters prefer x to y, a majority prefer y to z, and a majority prefer z to x. Unfortunately, this result is not due to a removable imperfection in the method of majority voting. The four conditions on social choice are mutually contradictory.

The philosophical and distributive implications of the paradox of social choice are still not clear. Certainly, there is no simple way out. I hope that others will take this paradox as a challenge rather than as a discouraging barrier.

REFERENCES

K. J. Arrow, (1951a) "An Extension of the Basic Theorems of Classical Welfare Economics," in J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley and Los Angeles 1951, 507–32.

———, (1951b) *Social Choice and Individual Values,* New York 1951.

———, "Le Rôle de Valeurs Boursières pour la Repartition la Meilleure des Risques," *Econometrie,* Colloques Internationaux du Centre National de la Recherche Scientifique, Vol. XI, 1953, 41–47.

———, (1963a) "Uncertainty and the Welfare Economics of Medical Care," *Amer. Econ. Rev.,* Dec. 1963, *53,* 941–73.

———, (1963b) *Social Choice and Individual Values,* New York 1963.

———, "Models of Racial Discrimination," in A. H. Pascal, ed., *Racial Discrimination in Economic Life,* Lexington 1972.

——— and G. Debreu, "Existence of an Equilibrium for a Competitive Economy," *Econometrica,* July 1954, *22,* 265–90.

——— and F. H. Hahn, *General Competitive Analysis,* San Francisco 1971.

A. Bergson, "A Reformulation of Certain Aspects of Welfare Economics," *Quart. J. Econ.,* Feb. 1938, *52,* 310–34.

K. G. Cassel, *Theoretische Sozialökonomie,* Erlangen 1918.

Marquis de Condorcet, *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix,* Paris 1785.

G. Debreu, "The Coefficient of Resource Utilization," *Econometrica,* July 1951, *19,* 273–92.

———, *Theory of Value,* New York 1959.

D. Gale, "General Equilibrium of Linear Models," unpublished 1957.

J. R. Hicks, *Value and Capital,* Oxford 1939.

——— and R. G. D. Allen, "A Reconsideration of the Theory of Value," *Economica,* Feb. 1934, *1,* 52–76, 196–219.

H. Hotelling, "Demand Functions with Limited Budgets," *Econometrica,* Jan. 1935, *3,* 66–78.

———, "The General Welfare in Relation to Problems of Taxation and Railway and Utility Rates," *Econometrica,* July 1938, *6,* 242–69.

S. Kakutani, "A Generalization of Brouwer's Fixed-Point Theorem," *Duke Mathematical J.,* 1941, *8,* 451–59.

J. M. Keynes, *The General Theory of Employment Interest and Money,* New York 1936.

T. C. Koopmans, "Analysis of Production as an Efficient Combination of Activities," in Koopmans, ed., *Activity Analysis of Allocation and Production,* New York 1951, ch. 3.

O. Lange, "Say's Law: A Restatement and Criticism," in Lange et al., eds., *Studies in Mathematical Economics and Econometrics,* Chicago 1942, 49–68.

A. Leijonhufvud, *On Keynesian Economics and the Economics of Keynes: A Study of Monetary Theory,* New York 1968.

L. McKenzie, "On Equilibrium in Graham's Model of World Trade and Other Competitive Systems," *Econometrica,* Apr. 1954, *22,* 147–61.

———, "On the Existence of General Equilibrium for a Competitive Market," *Econometrica,* Jan. 1959, *27,* 54–71.

———, "On the Existence of General Equilibrium: Some Corrections," *Econometrica,* Apr. 1961, *29,* 247–48.

H. Neisser, "Lohnhöhe und Beschäftigungsgrad im Marktgleichgewicht," *Weltwirtschaftliches Archiv,* 1932, *36,* 415–55.

J. von Neumann, "Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes," *Ergebnisse eines Mathematischen Kolloquiums,* 1939, *8,* 73–83.

——— and O. Morgenstern, *Theory of Games and Economic Behavior,* Princeton 1944.

V. Pareto, *Manuel d'Économie Politique,* Paris 1909.

R. Radner, "Competitive Equilibrium under Uncertainty," *Econometrica,* Jan. 1968, *36,* 31–58.

P. A. Samuelson, "A Note on the Pure Theory of Consumer Behavior," *Economica,* Feb. 1938, *4,* 61–71.

———, *Foundations of Economic Analysis,* Cambridge, Mass. 1947.

———, "Maximum Principles in Analytical Economics," in *Les Prix Nobel en 1970,* Stockholm 1971, 273–88.

K. Schlesinger, "Über die Produktionsgleichungen der ökonomische Wertlehre," *Ergebnisse eines Mathematischen Kolloquiums,* 1933–34, *6,* 10–11.

H. von Stackelberg, "Zwei Kritische bemerkungen zur preistheorie Gustav Cassels," *Zeitschrift für Nationalökonomie,* 1933, *4,* 456–472.

A. Wald, "Über die eindeutige positive Lösbarkeit der neuen Produktionsgleichungen," *Ergebnisse eines Mathematischen Kolloquiums,* 1933–34, *6,* 12–20.

———, "Über die Produktionsgleichungen der ökonomische Wertlehre," *Ergebnisse eines Mathematischen Kolloquiums,* 1934–35, *7,* 1–6.

———, "Über einige Gleichungssysteme der mathematischen Ökonomie," *Zeitschrift für Nationalökonomie,* 1936, *7,* 637–70.

Capital Importation and Sectoral Development: A Model Applied to Postwar Puerto Rico

By JOHN J. MINGO*

This paper represents an attempt to bridge two separate areas of inquiry. There exists a body of literature which analyzes capital movements in terms of the response of the world capital market to the growth, investment, and savings levels of open economies (see articles by George Borts (1964, 1968), Borts and Kenneth Kopecky, James Ingram, Harry Johnson, and Philip Neher). This literature has tended to ignore the sectoral aspects of industrial development, particularly the effects of internal labor mobility on the development of industrial sectors in capital importing regions. Inquiry into labor allocation and its impact on growth has generally been reserved for the analysis of closed dual economies (see articles by J. C. H. Fei and Gustav Ranis (1963, 1966), Dale Jorgenson, Arthur Lewis, and Michael Todaro). In these models, the migration of labor from a low-wage, low-productivity traditional agriculture sector to a high-productivity industrial sector becomes the key element in the development process. The dual economy literature, however, has generally retained the assumption that additions to the capital stock are financed solely from domestic savings and that physical capital may not be imported from outside the economy.

In this paper the dual economy is analyzed under the assumptions that it can trade on world commodity markets and borrow on world capital markets; and we examine the growth process as such an economy undergoes industrialization, that is, a reallocation of labor toward a more capital-intensive sector. The model presented will be tested against the experience of postwar Puerto Rico.

I. Closed and Open Models of Development

The output growth of the closed economy depends on technological change, savings, and the growth of the labor force. The closed dual economy is seen to develop through the transfer of labor from agriculture to industry. In some models, agricultural labor is physically redundant in the sense of having zero or negative marginal productivity; in others it transfers as a consequence of technological improvements which raise per capita agricultural output above demand. In either approach, the worker is equipped with machines when he transfers, such machines provided first by the savings of land owners and later by the savings of entrepreneurs.

Once in existence, the industrial sector of a closed economy grows at a rate which is constrained by the rate of savings in the economy and by the rate of technical progress in agriculture. Savings provide the necessary funds for capital accumulation, and the rate of technical progress in agriculture determines the rate at which farm labor can be released for work in the industrial sector.[1]

---

* Economist, Board of Governors, Federal Reserve System. I wish to thank an anonymous referee for comments made on an earlier draft.

In contrast, the output growth of the open economy depends on foreign markets for exports and on foreign savings. The demand for the economy's output is not constrained by the earnings of domestically owned factors of production. Foreign demand for exports will make it profitable to produce for a world market. Moreover, foreign savings permit workers to be equipped at a faster rate than reliance on domestic savings alone.

As in the closed, dual models, reallocation of labor toward a capital-intensive (hence, high labor productivity) industrial sector raises overall per capita output. However, as labor migrates to the capital-using sector, the marginal productivity of capital rises, the level of investment rises, and the open economy borrows more. Simultaneously, capital purchases (financed by foreign borrowing) contribute to high labor productivity and high wages in the industrial sector, which are necessary as an impetus to the reallocation of labor. Long-term borrowing and intersectoral labor migration, therefore, are closely related within the growth process.

The income or Gross National Product of the open economy is not equal to its output (Gross Domestic Product). The difference is the return flow of interest and dividends on borrowed capital, Net Factor Income (*NFI*). *NFI* is an element of the balance of payments on current account, and it grows over time with the country's indebtedness. A number of writers (see articles by Walter Salant, Evsey Domar, and David Avramovic and Ravi Gulhati) have considered the possibility that *NFI* might grow out of all proportion to a borrower's ability to repay, as measured by gross product or by exports. Such a problem does not arise in the theoretical framework employed here, since all monetary magnitudes grow at the same rate in dynamic equilibrium. In the model to be presented below, the ratio of *NFI* to *GNP* is one of the crucial dependent variables. Specifically, the time path of that ratio becomes important in explaining the interrelation between the borrowing-lending position of the dual region and the level of development.

In Section II, an open two-sector growth model is presented and the solution and stability of the system are considered in Appendix A. In Section III, the effects of key parametric changes are examined (some of which may be construed as government policy changes). In Section IV, data from postwar Puerto Rico are used to estimate the parameters of the system. Time paths for crucial dependent variables are then generated by computer simulation and compared with actually observed values. The use of the dynamic simulation is an integral part of the paper. First, it represents an attempt to fit the open, neoclassical growth model to a body of data representing a developing country. This allows a comparison of the predictive accuracy of the model with prior work, and leads to the conclusion that the open growth model is of some real use in explaining the development process. Second, the simulation may be used to show the effect on the paths of crucial variables of any given parametric change; for example, in Section IV we see the effect of a dramatic change in the rate of growth of labor on the borrowing-lending position of the region. Using multiple regression, for example, one cannot uncover a statistically significant relationship between the labor force growth rate and the borrowing-lending position.[2]

---

[1] Technical progress in agriculture serves a dual purpose in the closed model: unless per capita output (income) in agriculture exceeds the level of per capita food demand, there can be no consumption of industrial products. Industrial workers cannot be relied on to provide sufficient demand for their products, since part of their income is spent on agricultural products (food). See Jorgenson.

II. The Model

The model employed in this paper is a two-sector open growth model of a small country. The country sells its export commodity and buys its imports at perfectly elastic world prices, and borrows at a going world interest rate. The two sectors are both output sectors: a high-wage, industrial (or "advanced") sector, and a low-wage, agricultural (or "traditional") sector.[3] Physical capital inputs are solely imported. Relative labor supply between sectors is a function of relative wages; overall labor supply is treated as exogenous. Capital moves between the country (Puerto Rico) and the rest of the world (the United States) so as to equalize rates of return among various investment opportunities. That is, capital funds flow into or out of the region such that the ratio of the marginal revenue product of capital to its supply price equals the world rate of return:

$$r = \frac{MP_k \cdot P_x}{P_k} \tag{1}$$

where r is the going *U.S.* interest rate, P_k is the price of a unit of capital, MP_k is the marginal physical product of a unit of capital, P_x is the price of a unit of output in the capital-using sector.

A common currency exists between Puerto Rico and the United States, there are no interregional gold flows, and the demand for money is ignored. Therefore, the borrowing country, with a positive balance on capital account, experiences an offsetting negative balance on current account. For simplicity, the capital account consists of one type of debt instrument, a consol.

A country that has been a borrower in the past will make a payment in each time period on its outstanding debt instruments held abroad. This payment will be equal to the world interest rate multiplied by the value of outstanding debts. This payment is Net Factor Income (*NFI*), and it refers to net returns paid to foreign-owned capital. For simplicity, *NFI* does not include wages sent home from *labor* employed abroad.

Since the current account is the negative of the capital account, a country's current borrowing will be equal to its negative balance on current account at the present time. A country that has been a net borrower in the past will experience a *negative* Net Factor Income (*NFI*) in the present. Write *NFI* as,

$$D_t = r \int^{t} B_c(\tau) d\tau \tag{2}$$

where $B_c(\tau)$ is the balance on current account in period τ.

GNP in the country, therefore, is equal to *GDP* (gross domestic product valued at current market prices) *plus NFI* (which may be positive or negative). Clearly, for a country which has been a net borrower in the past, *NFI* is negative and, therefore, *GNP* is less than *GDP*. Write *GNP* as

$$Z = XP_x + YP_y + D \tag{3}$$

where XP_x is the dollar value of output in the X, or capital-intensive sector; YP_y is the dollar value of output in the labor-intensive sector; and D is net factor income.[4]

---

[2] One is likely to obtain insignificant regression coefficients for the model's parameters (when treated as time variables) for several reasons: The independent variables in the time-series may not vary much over time. Second, "significance" is often a question of taste (we reject a coefficient at the 10 percent level but accept it at the 5 percent level). Finally, in a large system, multicollinearity will obscure what would otherwise be significant coefficients.

[3] The model corresponds roughly to Jorgenson's third, or neoclassical, stage of development in that workers are paid their (positive) marginal products. Following Jorgenson, it is assumed that the marginal revenue product of industrial workers is higher than that of farm workers. The assumption is not crucial to the model, but can be used to ensure global stability (see Appendix A).

GNP may either be consumed or saved.

$$Z = C + S \tag{4}$$

and savings (including the government deficit) is a constant proportion of GNP.

$$S = \alpha Z \tag{5}$$

The industrial sector's good may either be consumed or exported (it is not a capital good).

$$X = X_c + X_e \tag{6}$$

where X_c is home consumption of the X good, X_e is physical exports.

The agricultural sector's good is wholly consumed at home; therefore,

$$C = X_cP_x + YP_y \tag{7}$$

where P_x and P_y are the prices of X and Y, respectively.

The demand for the Y good is further defined as being a constant proportion of GNP:[5]

$$YP_y = \eta Z \tag{8}$$

The balance of payments on current account equals exports minus imports plus net factor income,

$$B_c = X_eP_x - M\cdot P_k + D \tag{9}$$

where X_eP_x is the value of total exports, M is physical imports, P_k is the price of imports. Notice that in aggregate equilibrium $B_c =$ Savings $-$ Investment.

The production function in the X (capital-using) sector is a Cobb-Douglas with neutral technological change,

$$X = A\cdot L_x^{1-a}K^a \tag{10}$$

where A is the technological scalar, $(\dot{A}/A = \sigma)$, L_x is labor employed in the X sector, K is physical capital employed in the X sector.

Labor is the only variable factor in the production of Y,

$$Y = L_y\cdot b \tag{11}$$

where b is a (constant) technical scalar. (Note that b may be thought of as a combination of an invariant capital stock and a technical scalar, which may grow over time, if desired. Equation (11) is simple and the X sector is clearly the capital-intensive sector.)

Since the region is small, the world interest rate and the prices of exports and imports are given to it; world prices grow at constant exogenous rates:

$$r = \bar{r} \tag{12}$$

$$P_x = P_{x_0}e^{\rho t} \tag{13}$$

$$P_k = P_{k_0}e^{\gamma t} \tag{14}$$

Labor is paid its marginal product in both sectors.

$$w_x = (1 - a)XP_x/L_x \tag{15}$$

$$w_y = P_y\cdot b \tag{16}$$

where w_x and w_y are money wages in the X and Y sector, respectively.

Capital goods are solely imported, hence

---

[4] In Puerto Rico, GNP is measured at market prices while domestic output $(XP_x + YP_y)$ is valued at factor prices; it is the sum of payments to factors of production, determined by prices and production functions. To the extent that the government taxes the product of a firm (e.g., a sales tax), rather than income paid to factors of production, there is a wedge driven between the *market* value of gross output and the sum of payments to the factors of production. Equation (3) above should be rewritten as

$$Z = XP_x + YP_y + T + D \tag{3'}$$

where Z is the market value of income, XP_x and YP_y are valued at factor prices, T represents all indirect taxation of firms by the government *less* any direct subsidies made to firms by the government. This equation will be used in tests of the model against Puerto Rican data because produced output is *measured* at factor prices and includes taxes on profits but excludes commodity taxes. Also, indirect taxes are assumed to be a constant proportion of income, $T = \theta Z$.

[5] For simplicity, assume unitary income and price elasticities in (8). Also, η may fall over time (as relative demand for the Y good falls) without destroying the theoretical simplicity of the model.

additions to the capital stock are defined as equal to imports in physical units.[6]

$$\dot{K} \equiv M \tag{17}$$

Investment is defined as the value of net additions to the capital stock.[7]

$$I = \dot{K}P_k \tag{18}$$

The overall labor force is the sum of labor employed in the two sectors and grows at a constant exogenous rate,

$$L_x + L_y = L_0 e^{\lambda t} \tag{19}$$

Laborers offer their services to the two sectors according to the wage ratio,

$$L_x/L_y = \omega[w_x/w_y] \tag{20}$$

where ω is a constant. Equation (20) is a relative labor supply function, stating that the relative labor supply between sectors is linearly related to the sectoral money wage ratio (with a zero intercept). Certainly other choices might be appropriate. The main consideration is to allow labor to supply itself to the sectors in response to some economic motive such as a wage differential. To assume, for example, that the ratio of real or money wages remains constant over time is much less desirable because it supposes a solution in the sectoral labor market.[8] Note that by specifying a relative labor supply function there is no implication that wage rates would tend toward equality in the long run. The relevant tastes of labor in the two sectors could very well permit a wage differential to exist over time. Also I have found that equation (20) is not inconsistent with Puerto Rican data.

III. Solutions and the Effects of Parametric Changes[9]

A. *The Steady State*

In the steady state all monetary magnitudes grow at rate

$$g = \frac{\sigma + \rho - a\gamma}{1 - a} + \lambda$$

(a result proved in Appendix A). Industrial output in physical units grows at

---

[6] The assumption that imports consist solely of capital goods is retained so that the theoretical model is not unduly complicated. In reality, however, a region would need to spend some of its factors on producing inputs for the industrial process. To the extent the region can import these inputs, or import some foodstuffs whose production would otherwise occupy farm workers, more labor can be released for use in the industrial output sector. As this class of imports (other than physical capital) increased, per capita Gross Domestic Product would rise, but *NFI* would decline. That is, the balance on capital account would deteriorate—both to pay for the increased imports and because more workers available for employment in the X sector implies a higher marginal efficiency of investment there—hence, more capital funds inflows become necessary to maintain a rate of return in industry equal to the world rate of return. Indeed, elsewhere I have found that a) Puerto Rico achieved a secular decline in its relative need for domestic factors to produce industrial inputs (by importing those inputs), and that b) this decline resulted in accelerations in the growth of per capita income, in the migration of workers to the industrial sector, and in the growth of long-term regional indebtedness.

[7] Depreciation is excluded so that the country's borrowing in each period is the difference between net savings and net additions to the capital stock. Note that under the definitions employed, we obtain the following investment function by differentiation of equation (1) with respect to time:

$$I = \dot{K}P_k = \frac{a}{r} X P_x \left[\frac{\dot{X}}{X} + \rho - \gamma\right]$$

[8] In the closed, dual models (Jorgenson, Lewis), it is often assumed that the relative labor supply function has a zero slope. This assumption is inconsistent with actual observations that intersectoral wage differentials exist and often grow over time. Moreover, the assumption allows the industrial sector to expand without incurring the cost of a rising relative industrial wage. Hence, in Jorgenson's model, the asymptotic growth rate of the industrial labor force exceeds that of the agricultural labor force (p. 302), such that L_x/L_y rises without bound. Such a possibility does not exist in the present model: first, because the sectoral labor transfer can be accomplished only at the expense of a rising w_x/w_y; and second, because the rate of growth of investment (necessary to maintain a rate of return in the industrial sector equal to the world rate) declines toward an asymptote at which no further increases in the relative demand for industrial labor are forthcoming. See the discussion in Section III pertaining to Figure 1.

[9] In the presentation of these solutions, we ignore θ the excise tax rate discussed in fn. 4, although it is used in the empirical presentation in Section IV.

rate $k=g-\rho$. Also, the intersectoral labor ratio L_x/L_y reaches a stable asymptotic value such that labor grows at rate λ in each sector. Given these results, and utilizing the capital-market equation (1), we obtain the steady-state rate of growth of the industrial capital stock,

$$\dot{K}/K = k + (\rho - \gamma) \tag{21}$$

That is, the output-capital ratio in industry will remain constant provided there is no secular change in the terms of trade (between exports and imported capital goods). If the terms of trade are improving over time ($\rho>\gamma$), the rate of capital accumulation will have to exceed the rate of output growth in order to maintain a constant rate of return in the industrial sector (equal to the world rate of return).

The ratios D/Z (net factor income to *GNP*) and w_x/w_y have constant steady-state values. Specifically,

$$\left(\overline{\frac{XP_x}{Z}}\right) = \frac{g(1-\eta) - r\alpha}{g(1-a) + a\gamma} \tag{22}$$

(See this by dividing (A1) in Appendix A by Z and noting that all monetary variables grow at rate g in the steady state.)

Since the definition of *GNP* may be rewritten as $D/Z=1-\eta-XP_x/Z$ we obtain

$$\left(\frac{\overline{D}}{Z}\right) = \frac{r\alpha - a(g-\gamma)(1-\eta)}{g(1-a) + a\gamma} \tag{23}$$

Now, in order to derive the steady-state wage ratio, write the ratio of the total wage bills in each sector as

$$\frac{W_x}{W_y} = \frac{L_x w_x}{L_y w_y} = \frac{(1-a)XP_x}{YP_y} = \frac{(1-a)XP_x}{\eta Z} \tag{24}$$

and employing (22)

$$\frac{W_x}{W_y} = \frac{(1-a)}{\eta}\left[\frac{g(1-\eta) - r\alpha}{g(1-a) + a\gamma}\right] \tag{25}$$

Finally, since we assume $L_x/L_y=\omega(w_x/w_y)$, we may write

$$\frac{W_x}{W_y} = \left(\frac{w_x}{w_y}\right)^2 \omega$$

and therefore

$$\left(\overline{\frac{w_x}{w_y}}\right) = \left(\frac{(1-a)}{\eta\omega}\left[\frac{g(1-\eta) - r\alpha}{g(1-a) + a\gamma}\right]\right)^{1/2} \tag{26}$$

Hence, in the steady state, the sectoral wage ratio and the sectoral labor ratio are constant.[10]

Industrial output per worker grows at rate $g-\lambda$ in the steady state (see Appendix A), because of technical progress in the industrial sector and because of increases in the capital-labor ratio (brought about by improvements in the terms of trade and, hence, in the marginal efficiency of investment). Finally, overall *GNP* per worker also grows at rate $g-\lambda$ in the long run.

B. *Parametric Changes in the Steady State*

We shall examine the impact of four parameters (λ, α, ω, $\rho-\gamma$) on the steady-state values of three variables (D/Z, w_x/w_y, Z/L). Of the four parameters, three can be changed by domestic policy. The fourth ($\rho-\gamma$) is a condition which exists in world commodity markets (the rate of change of the terms of trade), and is beyond the control of the small region.

Effects on Borrowing-Lending:

$$\frac{\partial(\overline{D/Z})}{\partial\lambda} < 0 \qquad \frac{\partial(\overline{D/Z})}{\partial\alpha} > 0 \tag{27}$$

---

[10] Although the sectoral wage ratio is constant in the steady state, the output price ratio (P_x/P_y) may decline over time because of technical change and/or a rising capital-labor ratio in the industrial sector. Specifically, assuming no technical change in the domestic sector, P_y grows at rate $(\sigma/(1-a))+(a(\rho-\gamma)/(1-a))+\rho$, which is $(g-\lambda)$, the growth of *GNP* per worker in the long run.

$$\frac{\partial(\overline{D/Z})}{\partial\omega} = 0 \qquad \frac{\partial(\overline{D/Z})}{\partial(\rho-\gamma)} < 0$$

An increase in the labor force growth rate λ will increase the marginal efficiency of investment and induce the region to borrow more (D will turn more negative for a debtor region). An increase in the savings rate α, at any level of investment indicated by the capital market equation, implies less borrowing. An increase in ω, the labor supply coefficient, implies that labor is more "sensitive" to higher relative wages in industry, and will migrate to that sector more readily. This implies an increase in the marginal efficiency of investment in the industrial sector and leads to more borrowing. In the *long run*, however, the borrowing-lending position is not affected by ω. This is because *ceteris paribus* the borrowing-lending behavior of the region depends on the rate of growth of the industrial labor force—which grows at rate λ in the long run, no matter what value is attached to ω (i.e., no matter what are the long-run ratios of w_x/w_y and L_x/L_y). Finally, an increase in the terms of trade implies an increase in investment and in borrowing (hence, net factor income declines).

Effects on the Wage-Ratio:

$$(28) \qquad \frac{\partial(\overline{w_x/w_y})}{\partial\lambda} > 0 \qquad \frac{\partial(\overline{w_x/w_y})}{\partial\alpha} < 0$$

$$\frac{\partial(\overline{w_x/w_y})}{\partial\omega} < 0 \qquad \frac{\partial(\overline{w_x/w_y})}{\partial(\rho-\gamma)} > 0$$

An increase in the labor force growth rate implies a higher marginal efficiency of investment, more capital inflows into the industrial sector, and, hence, a higher relative money wage in that sector. An increase in the savings rate (at any level of investment and industrial output) means that the region borrows less, hence both D and Z are higher. Since demand for the domestic product is functionally related to *received* regional income, Z, the relative demand for domestic products rises and w_x/w_y declines. An increase in ω means that labor migrates to the industrial sector more readily, hence depressing the relative industrial wage. Finally, an increase in the terms of trade implies more capital inflows into the industrial sector, and, therefore, a higher relative industrial wage.

Effects on Income per Worker:

$$(29) \qquad \frac{\partial(\overline{Z/L})}{\partial\lambda} < 0 \qquad \frac{\partial(\overline{Z/L})}{\partial\alpha} > 0$$

$$\frac{\partial(\overline{Z/L})}{\partial\omega} > 0 \qquad \frac{\partial(\overline{Z/L})}{\partial(\rho-\gamma)} > 0$$

An increase in the rate of growth of labor does not change the time path of per capita output in the *industrial* sector! (Because any decline in the marginal productivity of labor is immediately offset by capital inflows aimed at maintaining a constant rate of return, see Appendix A.) However, the increased labor force implies an increase in the marginal efficiency of investment, more borrowing and a decline in net factor income (if NFI were negative, it would turn more negative). Therefore, *received* regional income per person must decline. An increase in the savings rate implies less borrowing at any investment level, a higher level of net factor income, and, hence, a greater level of received regional income. An increase in ω, the labor force's "willingness" to migrate, means that the relative size of the more productive industrial sector is greater at any moment, and therefore overall per capita output is greater. Finally, an increase in the terms of trade implies greater capital inflows and a higher level of per capita output and income.[11]

C. *The Allocation of Labor in the Short and the Long Run*

We can better understand how the open, dual model works by analyzing the sectoral allocation of labor L_x/L_y in any given time period, and then viewing the transition to the steady-state value.

The sectoral allocation of labor in the short run (i.e., at any time t) is determined by the intersection of the short-run relative labor demand function and the (stationary) relative labor supply function $[L_x/L_y=\omega(w_x/w_y)]$ in Figure 1 below. The short-run relative labor demand function may be derived by writing money wages in each sector at any time t:

$$w_x = (1-a)XP_x/L_x$$
$$w_y = YP_y/L_y = \eta Z/L_y$$

and then writing the money wage ratio as

$$\text{(30)}\qquad \frac{w_x}{w_y} = \frac{(1-a)\left(\dfrac{XP_x}{L_x}\right)_0 e^{\beta t}}{\eta\left(\dfrac{Z}{L}\right)\left[\dfrac{L_x}{L_y}+1\right]}$$

(In Appendix A it is shown that per capita industrial output XP_x/L_x grows at the constant rate β.) In (30) we see that the relative demand for labor (L_x/L_y) is inversely related to the wage ratio (w_x/w_y); that is, the relative labor demand function is downward sloped (as in Figure 1). The position of (30) in space, however, will change from time period to time period. Specifically, (30) will shift upward (downward) to its long-run stationary position if, in the short run, the rate of growth of (Z/L) is less than (greater than) β.

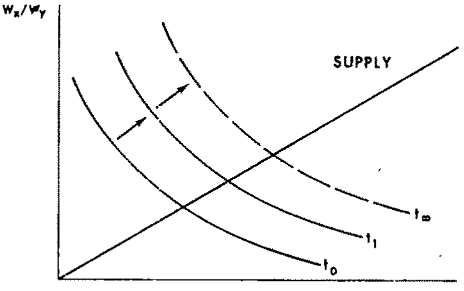


FIGURE 1. THE MOVEMENT OF THE RELATIVE LABOR DEMAND FUNCTION OVER TIME FOR A DEBTOR-DEVELOPING REGION

For a region which is undergoing industrialization, we would expect (30) to move *upward* to its long-run stationary position (such that L_x/L_y rises over time). Equation (30) *will* shift upward if, at the existing wage ratio, the rate of capital accumulation implied by the capital market equation (1) is such that there is an increase in the relative demand for industrial workers. Moreover, at the low savings rates associated with developing regions, such a rate of capital accumulation implies greater levels of regional indebtedness.[12] At some point in time, the upward movement of (30) will come to an end, for the relative industrial wage will have been bid high enough such that (at the existing wage ratio) the rate of capital accumulation dictated by the capital funds market implies a constant level of relative demand for industrial workers.[13]

---

[11] The policy maker will note that the four parameters do not all impact on the level of regional indebtedness and on per capita received income with results of similar sign. For example, if a small region experienced an improvement in its external terms of trade (say, through a trade agreement), per capita received income would rise, but at the expense of an increase in the relative level of regional indebtedness (which may not be politically desirable). As in the closed models, an increase in the savings rate (e.g., "forced savings" through taxation) or a decrease in the rate of growth of the labor force (through a population control program, say) would result in an increase in both per capita income and NFI—but again, the measures may be distasteful politically. A manpower program, on the other hand, aimed at increasing the willingness to migrate (say, by providing better market information) would increase per capita received income without affecting the percentage of the region's capital stock owned by foreigners.

[12] In fact, we may show unequivocally that for an existing debtor region, an upward movement of (30) entails a greater level of indebtedness. Begin by noting that at the existing sectoral labor ratio, (30) will shift upward if $Z^*<XP_x^*$ (where an asterisk indicates a percentage rate of change). Now, write *received* regional income as $(1-\eta)\,Z=XP_x+D$. It is clear that $Z^*<XP_x^*$ if $D^*<XP_x^*$. Now, the debtor will become *more* of a debtor if $I^*>S^*$. From the investment function posited in Section II, note that at any existing intersectoral

Since we *expect* a debtor-developing region to be characterized by an upward movement of equation (30) in Figure 1, we need to answer two crucial questions: First, do the dynamic properties of the model allow a feasible approach to a stable equilibrium consistent with the "comparative statics" treatment in Figure 1? (That is, *can* the system approach its stationary position such that, in the short run Z/L grows at a rate less than β, and w_x/w_y rises towards its asymptote?) Second, do the data show that for a typical debtor-developing region (Puerto Rico) per capita received income *has* grown slower than per capita industrial output, and that the relative industrial wage *has* risen over time?

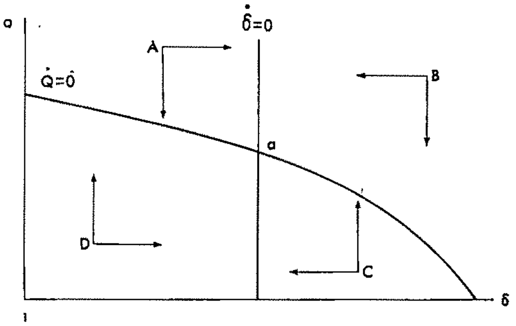


FIGURE 2

To answer the first question, define

$$\delta = w_x/w_y$$

and

$$Q = (Z/L)e^{-\beta t}$$

The model can then be represented by two simultaneous differential equations

$$\text{(A7)} \qquad \dot{\delta} = f(\delta)$$

$$\text{(A11)} \qquad \dot{Q} = h(\delta, Q)$$

(The explicit forms for (A7) and (A11) are derived in Appendix A.)

Assuming that the following parametric and initial conditions are met,

$$\delta_0 \geq 1; \qquad \left[\frac{r\alpha + \dfrac{\eta\omega a\gamma}{1-a}}{1-\eta-\eta\omega}\right] < g$$

the system may be described by the (inherently stable) phase diagram in Figure 2. Notice that the "comparative statics" treatment (in Figure 1) of an upward shift of the relative labor demand function is consistent with an economy which initially begins in Section A of the phase diagram. That is, the region will experience a rise in w_x/w_y and a decline in Q. (Since $Q=(Z/L)e^{-\beta t}$, this means Z/L grows at a rate less than β.)[14]

---

labor ratio $I^*=XP_x^*$. Also, the savings function ($S=\alpha Z$) implies $S^*=Z^*$. Therefore, if $Z^*<XP_x^*$ then $S^*<I^*$ and the region becomes more of a debtor.

[13] Even if the relative labor supply function were horizontal, a debtor-developing region would experience a short-run rightward shifting of the relative labor demand function, eventually coming to an end in the steady state. The definitions of w_x and w_y require that, at any sectoral wage ratio, the relative demand for the two labor types will remain unchanged only if the relative monetary demands for the X and Y goods remain unchanged; that is, only if $YP_y^*=Z^*=XP_x^*$. Such a condition is not the case in the model's short run, because Z and XP_x diverge by D (net factor income), and therefore do not automatically grow at the same rate. For example, for a debtor-developing region, we would expect that the investment and industrial growth (necessary to maintain a capital rate of return equal to the world rate) is quite high, such that $XP_x^*>Z^*$. In this case $XP_x^*>YP_y^*$ and (30) shifts rightward in the short run. The properties of the industrial production function dictate that the investment growth needed to satisfy the capital market equation declines over time (for a proof, see Appendix A). Therefore, XP_x^* declines to equality with Z^*, and in the steady state no further shifts in (30) occur. An upward sloped labor supply function, however, guarantees that the short-run increases in L_x/L_y are achieved only at the cost of permanent increases in the relative industrial wage—a result which is more consistent with reality than the assumption of a secularly maintained wage ratio.

[14] All observable points in the space represented by Figure 1 must lie *on* the (stationary) relative labor sup-

The second question is dealt with fully in Section IV. However, by way of introduction, it should be pointed out that Puerto Rico, in the period observed, behaved in precisely the way expected of a debtor-developing region. Specifically, per capita *received* income Z/L grew at a constant rate of .082, while per capita industrial output grew at the rate .095. Also the wage ratio w_x/w_y grew from 1.76 in 1948 to 3.36 in 1960. The sectoral labor ratio L_x/L_y grew from 1.75 to 3.38 in the same period. Finally, Puerto Rico became more of a debtor over that period, making net factor *payments* as a percentage of regional income of 1.1 percent in 1948 and 7.7 percent in 1960.

IV. Application of the Model to Postwar Puerto Rico

A. *Historical Summary*

Developments in the Puerto Rican economy suggest the applicability of the open dual model. Puerto Rico has experienced large capital inflows in the postwar years. Every year from 1947 to 1960, there was a negative balance on current account offset by a positive capital account. The capital account surplus was over 10 percent of net national product in 11 of the 14 years and over 43 percent of net investment. The latter ratio indicates the proportion of investment financed from abroad. As a result the net factor income outflow grew from 1.1 percent of net national income in 1950 to more than 10 percent in 1966. In dollar amounts, net factor income grew from $6.6 million in 1948 to $75.4 million in 1960.

Moreover, as the model suggests, there was a growth of industry relative to the agricultural sector of the economy. The labor ratio L_x/L_y grew from 1.75 in 1948 to 3.38 in 1960. During this period, industrial wages were rising relative to agricultural wages. Moreover, in 11 of the 13 successive paired years between 1947 and 1960, the value of industrial output grew more rapidly than gross domestic product. While all of this was going on, net national income per employed worker grew from $761 in 1947 to $2312 in 1960.

B. *A Test Through Simulation*

The model is not tested through traditional econometric methods. The small number of exogenous variables do not permit identification of all the structural parameters. Moreover, it is not possible with the present data to measure the rate of technological change in industry, an important element in the investment process. As an alternative, the model is simulated to determine its ability to track the motion of the dependent variables on the basis of assumed values of its key parameters. These assumed values are obtained by observation (over the years 1948–60) as suggested by the model itself. Any tracking errors will indicate parametric changes and/or changes in behavioral specification which can then be checked by further measurement in the period following 1960.

The assumed values of parameters necessary to simulate time paths for two crucial variables, Z/L, per capita national income and D/Z, the ratio of net factor income to national income are summarized in Table 1. Using these assumed parameter values and initial conditions, simulated paths for Z/L and D/Z are generated for the entire period 1948–70. Appendix B provides a description of methods used in the simulation. Table 2 below presents predicted and actual paths on a side-by-side comparison. Figures 3 and 4 represent

---

ply curve—because of the assumption of instantaneous adjustment in the capital funds market and in the labor market. However, the intersection of the relative supply function (20) and relative demand function (30) in Figure 1 *in the short run* will differ from the long-run intersection point—to the extent the region starts initially at a point other than point *a* in the phase diagram of Figure 2.

TABLE 1

Parameter or Initial Value	Value[a]	Explanation
g	.083	Steady-state rate of growth of all money magnitudes
λ	−.012	Rate of growth of labor force
α	.12	Savings rate
r	.17	World rate of return
η	.198	Share of income spent on food
b	.10	Output per farm worker
ω	.99	Coefficient of labor supply[b]
(Z/L_0)	\$770.	Initial *GNP* per capita[c]
$(XP_x/L_x)_0$	\$882.	Initial Industrial Output per worker

[a] Data sources are works by Stanley Friedlander, Ingram, Lloyd Reynolds and Peter Gregory, and George Stigler; as well as annual reports of the International Monetary Fund and the Puerto Rico Planning Board.

[b] Note that the parameter b is not needed since paths are given in dollar terms. Also the parameter ω was not used in the simulation. It was more convenient to generate paths for Z/L and D/Z using a fitted path of w_x/w_y, as described in Appendix B.

[c] Note that $(D/Z)_0$ is computed (predicted) given initial values for Z/L and XP_x/L_x.

the results graphically. In viewing the results it should be kept in mind that parameter estimates were made from the 1948–60 data. We can see from Figures 3 and 4 that the model, prior to 1960, generally underpredicts Z/L and D/Z. This discrepancy between the simulated paths and actual paths could be attributed

TABLE 2—PATHS OF *GNP* PER WORKER, Z/L, AND THE RATIO OF NET FACTOR INCOME TO *GNP*, D/Z, 1948–70

Year	(I) Actual Z/L	(II) Predicted Z/L	(III) Predicted Z/L'	(IV) Actual D/Z	(V) Predicted D/Z	(VI) Predicted D/Z'
1948	\$ 770	\$ 770	—	−.011	.029	—
1949	832	811	—	−.016	−.015	—
1950	885	865	—	−.028	−.056	—
1951	1018	932	—	−.025	−.082	—
1952	1284	1011	—	−.022	−.098	—
1953	1423	1102	—	−.028	−.098	—
1954	1466	1217	—	−.035	−.099	—
1955	1459	1336	—	−.046	−.099	—
1956	1519	1469	—	−.055	−.101	—
1957	1620	1618	—	−.059	−.097	—
1958	1891	1785	—	−.053	−.093	—
1959	2098	1970	—	−.061	−.088	—
1960	2312	2176	—	−.077	−.083	—
1961	2374	2404	2347	−.090	−.078	−.076
1962	2636	2656	2533	−.088	−.071	−.077
1963	2870	2935	2735	−.089	−.065	−.078
1964	2886	3244	2956	−.097	−.054	−.080
1965	3027	3585	3195	−.100	−.047	−.082
1966	3387	3961	3482	−.104	−.040	−.084
1967	3641	4375	3766	−.097	−.031	−.086
1968	3939	4875	4074	−.100	−.017	−.088
1969	4235	5337	4407	−.096	−.015	−.091
1970	4625	5892	4768	−.092	−.013	−.095

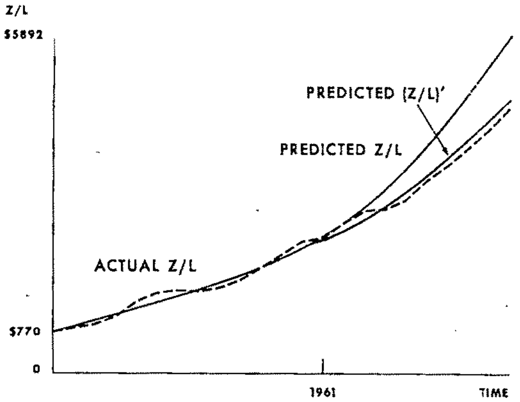


FIGURE 3

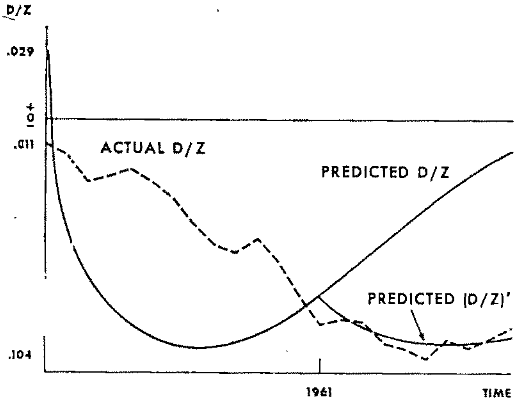


FIGURE 4

to a poor choice of parametric values. For example, higher values for α, the saving ratio, or r, the interest rate, or a lower value for λ, the growth of population, would lead to higher predicted values for Z/L, output/worker, and D/Z, the net factor income ratio. Of course, a breakdown in the accuracy of the simulation could also be attributed to inaccurate specification of behavioral equations.[15]

After 1960, the model generally predicts values for Z/L and D/Z that are too high. That is, the sign of the model's tracking error *reverses* itself after 1960. This result is almost certainly due to a dramatic parametric change that occurred after 1960. From 1948 to 1960 (the period from which original parameter estimates were made), the labor force *declined* at a constant rate of 1.2 percent losing almost 100,000 members. After 1960, the labor force *grew* at a constant positive rate of 3.1 percent.[16]

The effect of the more rapid growth of labor after 1960 was to inhibit the growth of per capita income and to induce more borrowing (that is, lower the D/Z ratio). Recall that the simulated paths of Z/L and D/Z were obtained by employing a constant parametric value for λ of -1.2 percent, a considerably lower value than the observed value of $+3.1$ percent in the post-1960 period. Since low values of λ theoretically raise the paths of Z/L and D/Z, it is reasonable to expect that simulated paths for Z/L and D/Z (employing a low estimate of λ) would be higher than actually observed Z/L and D/Z over the period 1960–70. Indeed, one may view this reversal of sign in the model's tracking error (see Figures 2 and 3) as a confirmation of the importance of the parameter λ in determining per capita income and the borrowing-lending position of the country. To further illustrate this point, paths for Z/L and D/Z are simulated, beginning at 1960, with $\lambda = .031$ (Table 2, columns III and VI, and designated in Figures 2 and 3 by $(Z/L)'$ and $(D/Z)'$.[17] Notice that the post-1960 predicted paths for

---

[15] As one example, the equation defining the demand for agricultural products could be respecified. First, the demand for farm products (in physical units) could be related proportionately to the overall population or labor force, that is, Y/L = constant. Second, we might assume that the elasticity of the dollar demand for farm products with respect to income is some constant, but not necessarily unity. Neither of these assumptions, however, is particularly consistent with Puerto Rican data.

[16] See Ingram, pp. 146–47, and Friedlander, p. 160, for a discussion of changes in the island's labor force.

[17] The predicted paths for Z/L and D/Z in columns III and IV of Table 2 reflect detected changes in λ *and* β in the post-1960 period. Specifically, actual β (see Appendix A for a definition) declined from .095 to .078 in the later (1960–70) period.

$(Z/L)'$ and $D/Z)'$ are much lower than the originally predicted paths, and more in line with the actually observed values of the variables.

An interesting feature of the model is that the (original) predicted path of D/Z (apart from any comparison with actual D/Z) falls over time but then begins to *rise* toward its asymptote. This phenomenon is consistent with the model, given the parameters which were chosen for prediction purposes. At first, D/Z falls (turns more negative) because the region is becoming more of a borrower and must make relatively more net factor payments. The region becomes more of a borrower because labor is being reallocated toward the industrial sector, leading to increases in the marginal productivity of capital (given a larger industrial labor force) and higher levels of investment. In the longer run, however, the wage ratio and sectoral labor ratio approach their asymptotes. At that point, the industrial labor force grows at the same rate as the overall labor force (the same rate as the agricultural labor force). Note, however, that the rate of growth of the overall labor force was negative during 1948–60 (the period from which parametric estimates were taken). Therefore, in the steady state, the industrial labor force would grow at a negative rate. This in turn means a lower marginal productivity of capital, lower investment levels, and less borrowing. The region makes fewer net factor payments to foreigners, and D/Z rises (turns less negative).

The remaining question to be answered is: How may we evaluate the accuracy and usefulness of the open, dual model? A good way would be to compare its predictions with those of another model. The most definitive earlier work on Puerto Rico was done by James Ingram. While Ingram's work concentrated on the regional payments aspect of the Puerto Rican economy, there exists a basis for comparing his model with the model presented here. Specifically, both models are capable of predicting capital-funds inflows into Puerto Rico.

Ingram uses standard one-period multiplier analysis to predict Puerto Rico's balance on current account (and, with no gold flows, its balance on capital account). In the second of his two models he assumes an exogenously given level of domestic investment and a fixed capital-income ratio. Given these (and initial values for all variables), he predicts changes in income, changes in consumption, changes in imports and changes in exports (changes in output which are not used for consumption or investment purposes are exported).[18]

To test the comparative predictive accuracy of Ingram's model we need a common predicted variable. Ingram predicts only the balance on current account. As equation (2) indicates, net factor payments are equal to

$$D = r \int B_c \tag{2}$$

where D is net factor payments, r is the world rate of return, and B_c is the balance on current account (negative of the balance on capital account). *Changes* in net factor payments, therefore, are equal to

$$\dot{D} = r \cdot B_c$$

That is, changes in net factor payments in any time period are equal to some going interest rate times each period's lending or borrowing (negative net factor incomes occur when B_c is negative; that is, the region is a net borrower). To compare Ingram's model with the two-sector open model, we may compare the two models' performance in predicting $\dot{D}$.[19]

---

[18] See Ingram, pp. 103–12.

[19] In the present model the predictor $\hat{\dot{D}}$, is the first difference of the term $(\widehat{D/Z}) \cdot (\widehat{Z/L}) \cdot (L)$. To compare

Using Henri Theil's coefficient of inequality to judge comparative predictive ability, we find that Ingram's model, in terms of its ability to predict D, yields a coefficient of inequality of .88; the two-sector open model yields a coefficient of .89. In effect this means that the models are very nearly equal in their ability (or inability) to predict what will happen. While this might indicate a tie, there is a strong argument for the superiority of the open two-sector model, because it requires *less* information for prediction purposes, than does Ingram's model. The neoclassical dual model does *not* take domestic investment as exogenous (it is a function of the marginal productivity of capital). It does not assume a fixed capital-output ratio (factors are substitutable in the production process). In this sense, the neoclassical model provides a more comprehensive explanation of reality, while losing little predictive usefulness.

In summary, the open dual model is helpful in explaining several phenomena connected with the development process in postwar Puerto Rico. First, a reallocation of labor toward the more productive industrial sector helped increase per capita output during the period observed. Second, the phenomenon of labor reallocation was intimately connected with the process of capital importation. The movement of labor toward the high wage sector coincided with increased investment and the overseas borrowing of capital funds. Similarly, capital importation and investment in the industrial sector contributed to high wages in that sector and acted as an impetus to labor reallocation. Third, during the period 1948 to 1960, the rate of growth of overall labor was negative due to migration to the mainland—contributing to the growth of per capita income. In the period 1960–70, however, migration to the mainland decreased and the rate of growth of overall labor increased, leading to a lower rate of growth of per capita income and more capital-funds borrowing. Had the rate of growth of overall labor remained negative in the 1960–70 period, the model predicts that Z/L would have been about \$1200 higher than actual 1970 per capita income. Moreover, Puerto Rico would have been less of a borrower, with predicted D/Z of $-.013$ instead of the actually observed $-.092$ in 1970 (see Table 2). Finally, the increases in the relative demand for industrial factors resulted in permanent increases in the relative industrial wage. For example, Reynolds and Gregory estimate that w_x/w_y almost doubled from 1948 to 1960. In short, the open two-sector model provides an explanation of the movements of key variables consistent with the theoretical relations between the borrowing-lending position of the developing country and the intersectoral composition of output and the labor force.

APPENDIX A

Stability and Equilibrium of the System

In this Appendix we show that w_x/w_y and L_x/L_y become constants in the steady state, and that XP_x, Z, YP_y and D all grow at the constant rate g in the steady state. Conditions for global stability of the system are also examined.

First, rewrite the relative labor supply function as $\epsilon = \omega\delta$ where $\epsilon = L_x/L_y$, $\delta = w_x/w_y$ and ω is a constant. Then,

$$Z = XP_x + YP_y + D$$

$$(1 - \eta)Z = XP_x + D$$

$$(1 - \eta)\dot{Z} = (\dot{XP_x}) + \dot{D}$$

Note that $\dot{D} = rB_c = r(S-I) = r\alpha Z - r\dot{K}P_k$ and from (1) in the text,

---

with Ingram's model, his $\dot{D} = r\cdot\hat{B}_c$ where $\hat{B}_c$ is his prediction of the current account, r is some fixed rate of return chosen as a scale to give Ingram's model the best fit between predicted $\hat{\dot{D}}$ and actual $\dot{D}$.

$$r\dot{K}P_K = a[\dot{X}P_x - XP_x\gamma]$$

so

$$\text{(A1)} \quad (1-\eta)\dot{Z} = (1-a)(\dot{X}P_x) + r\alpha Z + a\gamma XP_x$$

Continuing,

$$XP_x = \frac{XP_x}{L_x}(L - L_y)$$

$$= \frac{w_x}{1-a}\left(L - \frac{Y}{b}\right)$$

$$XP_x = \frac{w_x}{1-a}\left(L - \frac{YP_y}{bP_y}\right)$$

$$= \frac{w_x}{1-a}\left(L - \frac{\eta Z}{w_y}\right)$$

so

$$\text{(A2)} \qquad XP_x = \frac{1}{1-a}(w_xL - \eta Z\delta)$$

Now, given the capital market equation (1) and the production function for the X good,

$$w_x = w_{x_0}e^{\beta t}$$

where

$$\beta = \left(\frac{XP_x}{L_x}\right)^* = \frac{\sigma+\rho-a\gamma}{1-a}$$

where (*) indicates a constant rate of growth (see this from (10), (1), and (15)). Now, since $L^* = \lambda$, we may write

$$W = w_x \cdot L$$

$$W^* = \beta + \lambda = g$$

Equation (A2) above may then be rewritten as

$$\text{(A3)} \qquad XP_x = \frac{1}{1-a}(W - \eta Z\delta)$$

$$(\dot{X}P_x) = \frac{1}{1-a}(\dot{W} - \eta Z\dot{\delta} - \eta\delta\dot{Z})$$

Hence, substituting in (A1)

$$\text{(A4)} \quad (1-\eta)\dot{Z} = \dot{W} - \eta Z\dot{\delta} - \eta\delta\dot{Z} + r\alpha Z + \frac{a\gamma}{1-a}(W - \eta Z\delta)$$

We can convert (A4) into an equation in only $\dot{\delta}$ and δ. Call

$$U = W/Z$$

$$Z = W/U$$

then,

$$\dot{Z} = gZ - \frac{\dot{U}}{U}Z$$

Hence, substituting in (A4), we obtain

$$(1-\eta+\eta\delta)(gZ - U^*Z)$$

$$= \dot{W} - \eta\dot{\delta}Z + r\alpha Z + \frac{a\gamma}{1-a}W - \frac{a\gamma}{1-a}\eta\delta Z$$

or

$$\text{(A5)} \quad (1-\eta+\eta\delta)(g - U^*)$$

$$= \left[g + \frac{a\gamma}{1-a}\right]U - \eta\delta + r\alpha - \frac{a\gamma}{1-a}\eta\delta$$

Further note that

$$U = \eta(1 + L_x/L_y)(w_x/w_y)$$

or

$$U = \eta\delta + \eta\omega\delta^2$$

$$U = \eta\delta + 2\eta\omega\delta\delta$$

$$U^* = \frac{\eta\dot{\delta} + 2\eta\omega\delta\delta}{\eta\delta + \eta\omega\delta^2}$$

Therefore, we may rewrite (A5) as

$$\text{(A6)} \quad U^* - \frac{\eta\delta}{1-\eta+\eta\delta}$$

$$= g - \frac{\left[r\alpha + \left(g + \frac{a\gamma}{1-a}\right)U - \frac{a\gamma}{1-a}\eta\delta\right]}{1-\eta+\eta\delta}$$

Then, substituting in for U and U^*, and rearranging

$$\text{(A7)}\quad \dot{\delta}=\frac{r\alpha-g(1-\eta)+\eta\omega\left(g+\dfrac{a\gamma}{1-a}\right)\delta^2}{\dfrac{-(1-\eta+\eta\delta)(2\omega\delta+1)}{\delta(1+\omega\delta)}+\eta}$$

where (A7) is a non-linear differential equation with only one positive real root (steady-state value for δ),

$$\bar{\delta}=\left\{\frac{g(1-\eta)-r\alpha}{\eta\omega\left(g+\dfrac{a\gamma}{1-a}\right)}\right\}^{1/2}$$

To obtain an expression for the path of per capita income, write

$$z=Z/L=\frac{XP_x+YP_y+D}{L}$$

$$(1-\eta)\dot{z}=\frac{(X\dot{P}_x)}{L}-\frac{XP_x}{L}\cdot L^*+\frac{\dot{D}}{L}-\frac{D}{L}\cdot L^*$$

Recall that $\dot{D}=r\alpha Z-a[(X\dot{P}_x)-XP_x\gamma]$ that $L^*=\lambda$, and that $D=(1-\eta)Z-XP_x$; then,

$$\text{(A8)}\quad (1-\eta)\dot{z}=(1-a)\frac{(X\dot{P}_x)}{L}-[\lambda(1-\eta)-r\alpha]z+a\gamma\frac{XP_x}{L}$$

Also recall that $(X\dot{P}_x)=(1/1-a)[\dot{W}-\eta Z\dot{\delta}-\eta\delta\dot{Z}]$ so that (A8) becomes

$$(1-\eta)\dot{z}=\frac{\dot{W}}{L}-\eta z\dot{\delta}-\eta\delta\dot{z}-\eta\lambda\delta z-[\lambda(1-\eta)-r\alpha]z+\frac{a\gamma}{1-a}\left[\frac{W}{L}-\eta\delta z\right]$$

or

$$\text{(A9)}\quad (1-\eta+\eta\delta)\dot{z}=\left(g+\frac{a\gamma}{1-a}\right)w_{x_0}e^{\beta t}-\eta\dot{\delta}z-\eta\delta\left(\lambda+\frac{a\gamma}{1-a}\right)z-[\lambda(1-\eta)-r\alpha]z$$

Equation (A9) is used for the simulation of z, assuming a time path for δ which asymptotically stabilizes at $\delta=\bar{\delta}$. (See Appendix B.)[20] If we note that $\dot{\delta}=f(\delta)$ as in equation (A7), then we may represent (A7) and (A9) as two simultaneous differential equations of the form:

$$\text{(A7}')\quad \dot{\delta}=f(\delta)$$

$$\text{(A9}')\quad \dot{z}=g(z,f(\delta),t)$$

If we use the transformation $Q=ze^{-\beta t}$ where $\dot{Q}=e^{-\beta t}(\dot{z}-\beta z)$ then (A9) may be written:

$$\text{(A10)}\quad \dot{Q}=\frac{1}{1-\eta+\eta\delta}\cdot\left[\left(g+\frac{a\gamma}{1-a}\right)w_{x_0}-Q\left[\eta f(\delta)+\eta\delta\left(\lambda+\frac{a\gamma}{1-a}\right)+(\lambda(1-\eta)-r\alpha)\right]\right]-\beta Q$$

which is of the form

$$\text{(A11)}\quad \dot{Q}=h(\delta,Q)$$

Equations (A7′) and (A11) are two simultaneous differential equations in δ and Q. Since $f(\delta)$ is independent of $\dot{Q}$ and time, we may show conditions for global stability of (A7)

---

[20] If we assume a horizontal relative labor supply function, $\delta=\bar{\delta}$, the time path for received income in the region is derived simply: (A4) is replaced by

$$\text{(A4}')\quad (1-\eta)\dot{Z}=\dot{W}-\eta\delta\dot{Z}+r\alpha Z+\frac{a\gamma}{1-a}(W-\eta Z\delta)$$

using the definition of W, this becomes

$$\text{(A9}')\quad \dot{Z}=\phi_1 Z+\phi_2 W$$

which may be solved explicitly

$$\text{(9}'')\quad Z_t=\frac{\phi_2 W_0e^{gt}}{g-\phi_1}+R\ exp\ \{\phi_{1t}\}$$

$$\phi_1=\frac{r\alpha-a\gamma\eta\delta/(1-a)}{1-\eta+\eta\delta},\quad \phi_2=\frac{g+a\gamma/(1-a)}{1-\eta+n\delta}$$

where R is a positive, constant of integration. If $g>r\alpha-a\gamma\eta\delta/1-\eta+\eta\delta$ then $Z^*<g$, except in the steady state. Further, since $XP_x=(1/1-a)\,W_0e^{gt}-(\eta\delta/1-a)Z$; $XP_x^*>g$ (and $I^*>g$), except in the steady state. Hence, the relative labor demand curve shifts rightward and upward to its long-run position.

$$\text{(A12)}\quad \dot{\delta}=\left\{[r\alpha-g(1-\eta)]\delta+\omega[r\alpha-g(1-\eta)]\delta^2+\eta\omega\left(g+\frac{a\gamma}{1-a}\right)\delta^3+\eta\omega^2\left(g+\frac{a\gamma}{1-a}\right)\delta^4\right\}$$
$$\div\left\{-(1-\eta)2\omega\delta-(1-\eta)-\eta\omega\delta^2\right\}$$

by multiplying both numerator and denominator of (A7) by $\delta\,(1+\omega\delta)$ to obtain equation (A12). Since the denominator of (A12) is negative definite, $\dot{\delta}>0$ if the numerator is negative. As δ grows, (in the range $\delta\geq 1$), the coefficient of the term δ^4 will dominate such that $\dot{\delta}<0$; that is, δ will reach a stable asymptotic value. In the range $\delta<1$ there exists the possibility of an unstable equilibrium, but this is not crucial since we only need require that $(w_x/w_y)_0\geq 1$, that is, that the initial industrial wage is greater than or equal to the "backward" sector wage. A sufficient condition for global stability of (A12), therefore, is

$$\text{(A13)}\quad \frac{r\alpha+\dfrac{\eta\omega a\gamma}{1-a}}{1-\eta-\eta\omega}<g;\qquad \delta_0\geq 1$$

The term Q in (A10) will also reach a stable asymptotic value, if the conditions in (A13) are met. See this by noting that if $\delta=\bar{\delta}$, the phase curve represented by (A10) will be linear with a positive intercept and a negative slope; that is, inherently stable. If δ is approaching its asymptote from below (above), the function $\dot{Q}=h(\delta, Q)$ will rotate upward (downward) to its stationary long-run position. Finally, since in the steady state $\dot{\delta}=0$, see from (A9) that $(Z/L)^*=\beta$ and therefore $Z^*=\beta+\lambda=g$ (that is, all monetary magnitudes grow at rate g in the long run).

APPENDIX B

Simulation of the Economic System

Equation (A9) describes the path of z, per capita *GNP*,

$$\text{(A9)}\quad \dot{z}=\frac{1}{1-\eta+\eta\delta}\left[\left(g+\frac{a\gamma}{1-a}\right)w_{x_0}e^{\beta t}-\eta\dot{\delta}z-\eta\delta\left(\lambda+\frac{a\gamma}{1-a}\right)z\right.$$
$$\left.-(\lambda(1-\eta)-r\alpha)z\right]$$

for a known time path of δ, the wage ratio. This equation was used to simulate the time path of z. Using the parametric values and appropriate initial values listed in the text, a numerical simulation was carried out, using a fourth-order Runge-Kutta with ten increments for each unit of time. In order to simulate (A9), equation (A7) was assumed to be linear around its steady-state solution. Thus

$$\text{(A7}'\text{)}\qquad \dot{\delta}=A+B\delta$$

and

$$\text{(A7}''\text{)}\qquad \bar{\delta}_t=\bar{\delta}+Re^{Bt}$$

where R is a constant of integration and B is negative. A linearized path of δ was fitted to the actually observed path with the following coefficients

$$\text{(B1)}\qquad \delta=3.50-1.95e^{-.120t}$$

This equation fitted the data with a correlation coefficient of .93. Finally the ratio of *NFI* to *GNP* was calculated using

$$\text{(B2)}\quad D/Z=1-\eta-XP_x/Z$$

(as in the text); where

$$\frac{XP_x}{Z}=\left(\frac{XP_x}{L_x}\right)\left(\frac{L_x}{L}\right)\Big/\,\hat{z}$$

and

$$\frac{L_x}{L}=\frac{\omega\hat{\delta}}{1+\omega\hat{\delta}}\,;\quad \frac{XP_x}{L_x}=\left(\frac{XP_x}{L_x}\right)_0 e^{\beta t}$$

REFERENCES

D. Avramovic and R. Gulhati, *Debt Servicing Problems of Low Income Countries, 1956–1958,* Baltimore 1960.

G. H. Borts, "A Theory of Long-Run International Capital Movements," *J. Polit. Econ.*, Aug. 1964, *72*, 341–59.

———, "Growth and Capital Movements Among U.S. Regions in the Postwar Period," *Amer. Econ. Rev. Proc.*, May 1968, *58*, 155–61.

——— and K. J. Kopecky, "Capital Movements and Economic Growth in Developed Countries," in *Conference on International Mobility and Movement of Capital*, Universities–Nat. Bur. Econ. Res. conference series, New York 1972.

G. F. Carrier and C. E. Pearson, *Ordinary Differential Equations*, Waltham 1968, 183.

E. D. Domar, "The Effect of Foreign Investment on the Balance of Payments," *Amer. Econ. Rev.*, Dec. 1950, *40*, 805–26.

J. C. H. Fei and G. Ranis, "Innovation, Capital Accumulation, and Economic Development," *Amer. Econ. Rev.*, June 1963, *53*, 283–313.

——— and ———, "Agrarianism, Dualism, and Economic Development," in I. Adelman and E. Thorbecke, eds., *The Theory and Design of Economic Development*, Baltimore 1966.

S. L. Friedlander, *Labor Migration and Economic Growth, A Case Study of Puerto Rico*, Cambridge 1965.

J. R. Harris and M. P. Todaro, "Migration, Unemployment, and Development: A Two-Sector Analysis," *Amer. Econ. Rev.*, Mar. 1970, *60*, 126–42.

J. C. Ingram, *Regional Payments Mechanisms: The Case of Puerto Rico*, Chapel Hill 1962.

H. G. Johnson, "Equilibrium Growth in an International Economy," *Can. J. Econ.*, Nov. 1953, *19*, 478–500.

D. W. Jorgensen, "Surplus Agricultural Labor and the Development of a Dual Economy," *Oxford Econ. Pap.*, Nov. 1967, *19*, 288–312.

W. A. Lewis, "Economic Development with Unlimited Supplies of Labour," *Manchester Sch. Econ. Soc. Stud.*, May 1954, *22*, 139–91.

D. Mazumdar, "The Theory of Urban Unemployment in Less Developed Countries," unpublished paper.

J. J. Mingo, "Economic Dualism, Capital Imports, and the Demand for 'Backward'-Sector Products," *J. Develop. Stud.* forthcoming.

P. A. Neher, "A Neo-Classical Theory of International Capital Movements," unpublished doctoral dissertation, Brown Univ. 1966.

L. G. Reynolds and P. Gregory, *Wages, Productivity, and Industrialization in Puerto Rico*, Homewood 1965.

W. S. Salant, "The Domestic Effects of Capital Exports Under the Point Four Program," *Amer. Econ. Rev. Proc.*, May 1950, *40*, 495–510.

G. J. Stigler, *Capital Rates of Return in Manufacturing Industries*, Princeton 1968.

H. Theil, *Applied Economic Forecasting*, Amsterdam 1965.

M. P. Todaro, "A Model of Labor Migration and Urban Unemployment in Less Developed Countries," *Amer. Econ. Rev.*, Mar. 1969, *59*, 138–48.

International Monetary Fund, *Balance of Payments Yearbook*, 1947–1970.

Puerto Rico Planning Board, Bureau of Economics and Statistics, *Net Income and Gross Product*, 1960, 1961–62.

———, *Puerto Rico Statistical Yearbook*, 1947–1970.

The Political Economy of the Rent-Seeking Society

By ANNE O. KRUEGER*

In many market-oriented economies, government restrictions upon economic activity are pervasive facts of life. These restrictions give rise to rents of a variety of forms, and people often compete for the rents. Sometimes, such competition is perfectly legal. In other instances, rent seeking takes other forms, such as bribery, corruption, smuggling, and black markets.

It is the purpose of this paper to show some of the ways in which rent seeking is competitive, and to develop a simple model of competitive rent seeking for the important case when rents originate from quantitative restrictions upon international trade. In such a case 1) competitive rent seeking leads to the operation of the economy inside its transformation curve; 2) the welfare loss associated with quantitative restrictions is unequivocally greater than the loss from the tariff equivalent of those quantitative restrictions; and 3) competitive rent seeking results in a divergence between the private and social costs of certain activities. Although the analysis is general, the model has particular applicability for developing countries, where government interventions are frequently all-embracing.

A preliminary section of the paper is concerned with the competitive nature of rent seeking and the quantitative importance of rents for two countries, India and Turkey. In the second section, a formal model of rent seeking under quantitative restrictions on trade is developed and the propositions indicated above are established. A final section outlines some other forms of rent seeking and suggests some implications of the analysis.

I. Competitive Rent Seeking

A. *Means of Competition*

When quantitative restrictions are imposed upon and effectively constrain imports, an import license is a valuable commodity. It is well known that under some circumstances, one can estimate the tariff equivalents of a set of quantitative restrictions and analyze the effects of those restrictions in the same manner as one would the tariff equivalents. In other circumstances, the resource-allocational effects of import licensing will vary, depending upon who receives the license.[1]

It has always been recognized that there are *some* costs associated with licensing: paperwork, the time spent by entrepreneurs in obtaining their licenses, the cost of the administrative apparatus necessary to issue licenses, and so on. Here, the argument is carried one step further: in many circumstances resources are devoted to competing for those licenses.

The consequences of that rent seeking are examined below. First, however, it will be argued that rent-seeking activities are often competitive and resources are devoted to competing for rents. It is difficult, if not impossible, to find empirically observable measures of the degree to which rent seeking is competitive. Instead, some mechanisms under which rent seeking is almost certain to be competitive are examined. Then other cases are considered in which it is less obvious, but perhaps equally plausible, that competition results.

---

* Professor of economics, University of Minnesota. I am indebted to James M. Henderson for invaluable advice and discussion on successive drafts. Jagdish Bhagwati and John C. Hause made helpful comments on earlier drafts of this paper.

[1] This phenomenon is explored in detail in Bhagwati and Krueger.

Consider first the results of an import-licensing mechanism when licenses for imports of intermediate goods are allocated in proportion to firms' capacities. That system is frequently used, and has been analyzed for the Indian case by Jagdish Bhagwati and Padma Desai. When licenses are allocated in proportion to firms' capacities, investment in additional physical plant confers upon the investor a higher expected receipt of import licenses. Even with initial excess capacity (due to quantitative restrictions upon imports of intermediate goods), a rational entrepreneur may still expand his plant if the expected gains from the additional import licenses he will receive, divided by the cost of the investment, equal the returns on investment in other activities.[2] This behavior could be perfectly rational even if, for all entrepreneurs, the total number of import licenses will remain fixed. In fact, if imports are held constant as domestic income grows, one would expect the domestic value of a constant quantity of imports to increase over time, and hence installed capacity would increase while output remained constant. By investing in additional capacity, entrepreneurs devote resources to compete for import licenses.

A second sort of licensing mechanism frequently found in developing countries is used for imports of consumer goods. There, licenses are allocated *pro rata* in proportion to the applications for those licenses from importers-wholesalers. Entry is generally free into importing-wholesaling, and firms usually have U-shaped cost curves. The result is a larger-than-optimal number of firms, operating on the downward sloping portion of their cost curves, yet earning a "normal" rate of return. Each importer-wholesaler receives fewer imports than he would buy at existing prices in the absence of licensing, but realizes a sufficient return on those licenses he does receive to make it profitable to stay in business. In this case, competition for rents occurs through entry into the industry with smaller-than-optimally sized firms, and resources are used in that the same volume of imports could be efficiently distributed with fewer inputs if firms were of optimal size.

A third sort of licensing mechanism is less systematic in that government officials decide on license allocations. Competition occurs to some extent through both mechanisms already mentioned as businessmen base their decisions on expected values. But, in addition, competition can also occur through allocating resources to influencing the probability, or expected size, of license allocations. Some means of influencing the expected allocation—trips to the capital city, locating the firm in the capital, and so on—are straightforward. Others, including bribery, hiring relatives of officials or employing the officials themselves upon retirement, are less so. In the former case, competition occurs through choice of location, expenditure of resources upon travel, and so on. In the latter case, government officials themselves receive part of the rents.

Bribery has often been treated as a transfer payment. However, there is competition for government jobs and it is reasonable to believe that expected total remuneration is the relevant decision variable for persons deciding upon careers. Generally, entry into government service requires above-average educational attainments. The human capital literature provides evidence that choices as to how much to invest in human capital are strongly influenced by rates of return upon the investment. For a given level of educational attainment, one would expect the rate of return to be approximately equated among various lines of endeavor. Thus, if there appear to be high official-plus-unofficial incomes accruing to government officials and higher education is a prerequisite for seeking a government job, more individuals will invest in higher education. It is not necessary that government officials earn the same total income as other college graduates. All that is necessary is that there is an excess supply of persons seeking government employment, or that highly educated persons make sustained efforts to enter government services. Competition takes place through attaining the appropriate credentials for entry into government service and through accepting unemployment while making efforts to obtain appointments. Efforts to influence those in charge of making appointments, of course, just carry the argument one step further back.

---

[2] Note that: 1) one would expect to find greater excess capacity in those industries where rents are higher; and 2) within an industry, more efficient firms will have greater excess capacity than less efficient firms, since the return on a given amount of investment will be higher with greater efficiency.

To argue that competition for entry into government service is, in part, a competition for rents does not imply that all government servants accept bribes nor that they would leave government service in their absence. Successful competitors for government jobs might experience large windfall gains even at their official salaries. However, if the possibility of those gains induces others to expend time, energy, and resources in seeking entry into government services, the activity is competitive for present purposes.

In all these license-allocation cases, there are means, legal and illegal, for competing for rents. If individuals choose their activities on the basis of expected returns, rates of return on alternative activities will be equated and, in that sense, markets will be competitive.[3] In most cases, people do not perceive themselves to be rent seekers and, generally speaking, individuals and firms do not specialize in rent seeking. Rather, rent seeking is one part of an economic activity, such as distribution or production, and part of the firm's resources are devoted to the activity (including, of course, the hiring of expediters). The fact that rent seeking and other economic activities are not generally conducted by separate economic entities provides the motivation for the form of the model developed below.

B. *Are Rents Quantitatively Important?*

Granted that rent seeking may be highly competitive, the question remains whether rents are important. Data from two countries, India and Turkey, suggest that they are. Gunnar Myrdal believes India may " . . . on the balance, be judged to have somewhat less corruption than any other country in South Asia" (p. 943). Nonetheless, it is generally believed that "corruption" has been increasing, and that much of the blame lies with the proliferation of economic controls following independence.[4]

Table 1 presents crude estimates, based on fairly conservative assumptions of the value of rents of all sorts in 1964. One important source of rents—investment licensing—is not included for lack of any valid basis on which to estimate its value. Many smaller controls are also excluded. Nonetheless, it is apparent from Table 1 that

---

[3] It may be objected that illegal means of competition may be sufficiently distasteful that perfect competition will not result. Three comments are called for. First, it requires only that enough people at the margin do not incur disutility from engaging in these activities. Second, most lines of economic activity in many countries cannot be entered without some rent-seeking activity. Third, risks of detection (especially when bribery is expected) and the value judgments associated with illegal activities differ from society to society. See Ronald Wraith and Edgar Simpkins.

[4] Santhanam Committee, pp. 7–8.

TABLE 1—ESTIMATES OF VALUE OF RENTS: INDIA, 1964

Source of Rent	Amount of Rent (Rs. million)
Public investment	365
Imports	10,271
Controlled commodities	3,000
Credit rationing	407
Railways	602
Total	14,645

Sources:

1) Public investment: The Santhanam Committee, pp. 11–12, placed the loss in public investment at *at least* 5 percent of investment. That figure was multiplied by the average annual public investment in the *Third Five Year Plan.*

2) Imports: The Santhanam Committee, p. 18, stated that import licenses were worth 100 to 500 percent of their face value. Seventy-five percent of the value of 1964 imports was used here as a conservative estimate.

3) Controlled commodities: These commodities include steel, cement, coal, passenger cars, scooters, food, and other price—and/or distribution-controlled commodities, as well as foreign exchange used for illegal imports and other unrecorded transactions. The figure is the lower bound estimate given by John Monteiro, p. 60. Monteiro puts the upper bound estimate at Rs. 30,000 billion, although he rejects the figure on the (dubious) ground that notes in circulation are less than that sum.

4) Credit rationing: The bank rate in 1964 was 6 percent; Rs. 20.3 billion of loans were outstanding. It is assumed that *at least* an 8 percent interest rate would have been required to clear the market, and that 3 percent of bank loans outstanding would be equivalent to the present value of new loans at 5 percent. Data source: Reserve Bank of India, Tables 534 and 554.

5) Railways: Monteiro, p. 45, cites commissions of 20 percent on railway purchases, and extra-official fees of Rs. 0.15 per wagon and Rs. 1.4 per 100 maunds loaded. These figures were multiplied by the 1964 traffic volume; 203 million tons of revenue-paying traffic originated in that year. Third plan expenditure on railroads was Rs. 13,260 million. There were 350,000 railroad goods wagons in 1964–65. If a wagon was loaded once a week, there were 17,500,000 wagons of freight. At Rs. 0.15 per load, this would be Rs. 2.6 million; 100 maunds equal 8,228 pounds so at 1.4 Rs. per 100 maunds, Rs. 69 million changed hands; if one-fifth of railroad expenditures were made in 1964–65, Rs. 2652 million was spent in 1964; at 20 percent, this would be Rs. 530 million, for a total of Rs. 602 million.

import licenses provided the largest source of rents. The total value of rents of Rs. 14.6 billion contrasts with Indian national income of Rs. 201 billion in 1964. At 7.3 percent of national income, rents must be judged large relative to India's problems in attempting to raise her savings rate.

For Turkey, excellent detailed estimates of the value of import licenses in 1968 are available.[5] Data on the c.i.f. prices of individual imports, their landed cost (c.i.f. price plus all duties, taxes, and landing charges), and wholesale prices were collected for a sizeable sample of commodities representing about 10 percent of total imports in 1968. The c.i.f. value of imports in the sample was TL 547 million and the landed cost of the imports was TL 1,443 million. The value at the wholesale level of these same imports was TL 3,568 million. Of course, wholesalers incur some handling, storage, and transport costs. The question, therefore, is the amount that can be attributed to normal wholesaling costs. If one assumes that a 50 percent markup would be adequate, then the value of import licenses was TL 1,404 million, or almost three times the c.i.f. value of imports. Imports in 1968 were recorded (c.i.f.) as 6 percent of national income. On the basis of Aker's data, this would imply that rents from import licenses in Turkey in 1968 were about 15 percent of *GNP*.

Both the Indian and the Turkish estimates are necessarily somewhat rough. But they clearly indicate that the value of import licenses to the recipients was sizeable. Since means were available of competing for the licenses, it would be surprising if competition did not occur for prizes that large. We turn, therefore, to an examination of the consequences of competitive rent seeking.

---

[5] I am indebted to Ahmet Aker of Robert College who kindly made his data available to me. Details and a description of the data can be found in my forthcoming book.

II. The Effects of Competitive Rent Seeking

The major proposition of this paper is that competitive rent seeking for import licenses entails a welfare cost in addition to the welfare cost that would be incurred if the same level of imports were achieved through tariffs. The effects of tariffs upon production, trade, and welfare are well known, and attention is focussed here upon the additional cost of competitive rent seeking. A simple model is used to develop the argument. Initially, free trade is assumed. Then, a tariff or equivalent import restriction is introduced. Finally, an equal import restriction with competitive rent seeking is examined.

A. *The Basic Model*

Two commodities are consumed by the country under investigation: food and consumption goods. Food is produced domestically and exported. Consumption goods are imported. Distribution is a productive activity whereby food is purchased from the agricultural sector, exported, and the proceeds are used to import consumption goods which are sold in the domestic market. Labor is assumed to be the only domestic factor of production.[6] It is assumed that the country under consideration is small and cannot affect its international terms of trade. Physical units are selected so that the fixed international prices of both goods are unity.

The agricultural production function is

$$A = A(L_A) \qquad A' > 0,\ A'' < 0 \tag{1}$$

where A is the output of food and L_A is the quantity of labor employed in agriculture. The sign of the second derivative reflects a diminishing marginal physical product of labor in agriculture, due, presumably, to fixity in the supply of land.

The level of distribution output, D, is defined to equal the level of consumption-goods imports, M:

$$D = M \tag{2}$$

One unit of distributive services entails exchanging one unit of imports for food with the agricultural sector at the domestic terms of trade, and exporting the food in exchange for imports at the international terms of trade. Constant returns to scale are assumed for the distribution activity; one unit of distribution requires k units of labor. Total labor employed in distribution, L_D, is

$$L_D = kD \tag{3}$$

A distribution charge of p_D per unit is added to the international price of imports:

$$p_M = 1 + p_D \tag{4}$$

where p_M is the domestic price of imports. The domestic price of food is assumed to equal its unit international price.[7]

Society's demand for imports depends upon the domestic price of imports and total income generated in agriculture:[8]

$$M = M(p_M, A) \tag{5}$$

where $\partial M/\partial p_M < 0$ and $\partial M/\partial A > 0$. Demand decreases with increases in the price of imports, and increases with increases in agricultural output (income). Equation (5) is derived from micro utility maximization with the assumption that farmers, distributors, and rent seekers all have the same consumption behavior. Domestic food consumption, F, is simply the quantity not exported:

---

[6] Labor could be regarded as a composite domestic factor of production. Extensions to two or more factors would complicate the analysis, but would not alter its basic results.

[7] These assumptions establish a domestic numeraire. The real analysis would be unaffected by proportional changes in the domestic prices.

[8] Food and imports are consumed. But, by choice of food as the numeraire (see equation (6)) and the assumed constancy of international prices, agricultural output serves as a measure of income.

$$F = A - M \tag{6}$$

Since the fixed international terms of trade equal unity, food exports equal consumption goods imports.

Finally, it is assumed that the economy under consideration has a fixed labor supply, $\bar{L}$:

$$\bar{L} = L_A + L_D + L_R \tag{7}$$

where L_R is the quantity of labor engaged in rent seeking.

B. *Free Trade*

Under free trade, there is free entry into both agriculture and distribution and competition equates the wage in the two activities:

$$A' = p_D/k \tag{8}$$

Equations (1) to (8) constitute the free-trade system. These eight equations contain the eight variables A, M, D, F, L_A, L_D, p_M, and p_D. Since there is no rent seeking under free trade, $L_R \equiv 0$.

It is easily established that free trade is optimal in the sense that the domestic price ratio under free trade equals the marginal rate of transformation between food consumption and imports. The consumption possibility locus is obtained by substituting into (6) from (1) and (7)

$$F = A(\bar{L} - kM) - M$$

The locus has a marginal rate of transformation greater than one:

$$\frac{-dF}{dM} = kA' + 1 > 1 \tag{9}$$

which reflects the positive distribution cost of substituting imports for food consumption. The locus is concave:

$$\frac{d^2F}{dM^2} = k^2A'' < 0$$

since $A'' < 0$, which follows from diminishing returns in food production. Substituting from (8) into (9),

$$\frac{-dF}{dM} = 1 + p_D$$

which establishes the aforementioned equality.

A free-trade solution is depicted in Figure 1. Domestic food consumption and import consumption are measured along OF and OM, respectively. The consumption possibility locus is $\hat{F}\hat{M}$. At the point $\hat{F}$ no imports are consumed and hence there is no distribution. If distribution were costless, society could choose its consumption point from the line $\hat{F}A$. However, to consume one unit of import requires exchanging one unit of food *and* withdrawing k workers from agriculture to provide the requisite distributive services. With diminishing marginal product of labor in agriculture, the cost of additional imports in terms of foregone food production rises. Thus, the price of distribution, and hence the domestic price of imports, increases in moving northwest from $\hat{F}$. The consump-

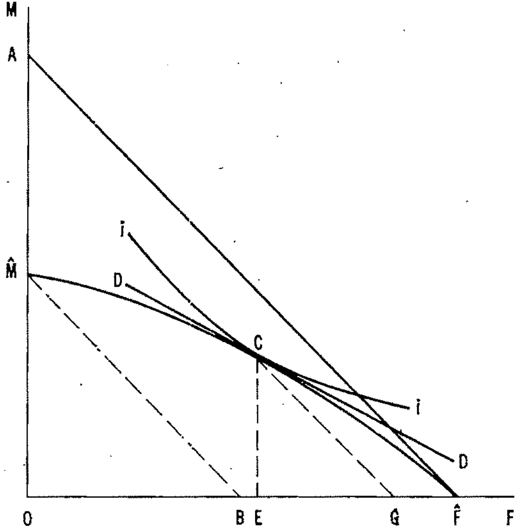


Figure 1. Free Trade

tion point $\hat{M}$ has OB food exchanged for $O\hat{M}$ of imports. The distance $\hat{F}B$ is the agricultural output foregone to distribute $O\hat{M}$ imports.

If society's preferences are given by the indifference curve ii, point C is optimal. The price of distribution is reflected in the difference between the slope of $\hat{F}A$ and the slope of DD at C. At the point C, OG food would be produced, with EG $(=EC)$ exported, and the rest domestically consumed.

C. *A Tariff or an Import Restriction Without Rent Seeking*

Consider now a case in which there is a restriction upon the quantity of imports

$$M = \overline{M} \tag{10}$$

where $\overline{M}$ is less than the import quantity that would be realized under free trade. Since entry into distribution is now limited, the competitive wage equality (8) will no longer hold. The relevant system contains (1) to (7) and (10). The variables are the same as in the free-trade case and again $L_R = 0$. The system may be solved sequentially: given (10), D follows from (2), L_D from (3), L_A from (7), A from (1), F from (6), p_M from (5), and p_D from (4). Since equations (1), (6), and (7) remain intact, the solution for this case is also on the consumption possibility locus.

It is useful to establish the directions of change for the variables following a switch from free trade to import restriction. The reduced import level will reduce the labor employed in distribution and increase the labor force in agriculture. Diminishing returns will reduce the agricultural wage. The domestic price of imports, the distributive margin, and the wage of distributors will increase. Distributors will earn a rent in the sense that their wage will exceed the wage of those engaged in agriculture.

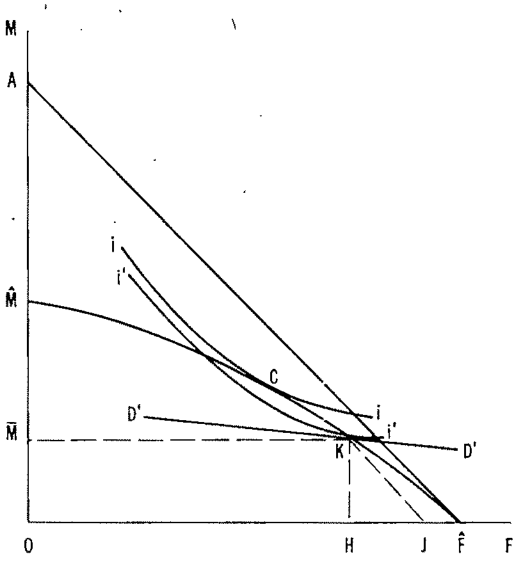


FIGURE 2. IMPORT RESTRICTION WITHOUT RENT SEEKING

In the absence of rent seeking, a tariff and a quantitative restriction are equivalent[9] aside from the resultant income distribution. Under a quantitative restriction the distributive wage is higher than the agricultural. If instead there were an equivalent tariff with redistribution of the proceeds, the marginal product of labor in agriculture would be unchanged, but agricultural workers would benefit by the amount of tariff proceeds redistributed to them whereas traders' income would be lower. Since the allocation of labor under a tariff and quantitative restriction without rent seeking is the same and domestic prices are the same, the only difference between the two situations lies in income distribution.

The solution under a quantitative restriction is illustrated in Figure 2, where $\hat{F}\hat{M}$ is again the consumption possibility locus and C the free-trade solution. With a quantitative restriction on imports in the amount $O\overline{M}$, the domestic prices of imports, and hence of distribution, rise from free trade to import restriction. Food output (OJ) and domestic consumption of food increase, and exports decline to HJ ($=OM$). The indifference curve $i'i'$ lies below ii (and the point C), and the welfare loss may be described by the consumption and production cost measure given by Harry Johnson.

---

[9] The change in the price of the import from the free-trade solution is the tariff equivalent of the quantitative restriction described here.

The wage rate in distribution unequivocally rises for a movement from free trade to a quantitative restriction. The total income of distributors will increase, decrease, or remain unchanged depending upon whether the proportionate increase in p_D is greater than, less than, or equal to the absolute value of the proportionate decrease of imports. For the moment, let p_D, p_M, and M represent free-trade solution values, and let p_D^*, p_M^*, and $\overline{M}$ represent import-restriction solution values. The total arc elasticity of demand for imports for the interval under consideration, η, is

$$\eta = \frac{-(\overline{M} - M)}{\overline{M} + M} \cdot \frac{p_M^* + p_M}{p_M^* - p_M} \tag{11}$$

Total expenditures on imports will increase, decrease, or remain unchanged as η is less than one, greater than one, or equal to one. The total income of distributors will increase if

$$p_D^* \overline{M} > p_D M$$

Multiplying both sides of this inequality by $(p_M^* + p_M)/(p_M^* - p_M)$, substituting from (11), and using (4),

$$1 + 2/(p_D^* + p_D) > \eta \tag{12}$$

Hence, distributors' total income can increase even if the demand for imports is price elastic.[10] The smaller is the free-trade distributive markup, the more likely it is that the distributors' total income will increase with a curtailment of imports. The reason is that an increase in the domestic price of imports results in a proportionately greater increase in the price of distribution.

D. *An Import Restriction with Competitive Rent Seeking*

In the import-restriction model just presented, the wage in distribution p_D/k exceeds the wage in agriculture A'. Under this circumstance, it would be surprising if people did not endeavor to enter distribution in response to its higher return. Resources can be devoted to rent seeking in all the ways indicated in Section IA. This rent-seeking activity can be specified in a number of different ways. A simple and intuitively plausible specification is that people will seek distributive rents until the average wage in distribution and rent seeking equals the agricultural wage:[11]

$$A' = \frac{p_D \overline{M}}{L_D + L_R} \tag{13}$$

One can regard all distributors and rent seekers as being partially engaged in each activity or one can think of rent seekers as entering in the expectation of receiving import licenses. In the latter case, the final solution classifies the successful seekers in L_D and the unsuccessful ones in L_R. Equation (13) implies risk neutrality in this circumstance.

The model for import restriction with rent seeking contains the same equations,

---

[10] Proof of (12) uses the step that $p_D^* \overline{M} > p_D M$ implies $(p_D^* - p_D)/(p_D^* + p_D) > -(\overline{M} - M)/(\overline{M} + M)$. Note that in the continuous case, (12) reduces to $1 + 1/p_D > \eta$.

[11] As an alternative, the distributive production function (3) can be altered to treat all persons competing for import licenses as distributors so that L_D also encompasses L_R and $A' = p_D \overline{M}/L_D$. Another alternative is to introduce a rent-seeking activity distinct from distribution with a wage determined from total rents $(p_D - A'k)\overline{M}/L_R$, and require that this wage equal the wages in distribution and agriculture. These specifications give results equivalent to those that follow from (13).

(1) to (7) and (10), and the same variables as the model for import restrictions without rent seeking. In addition, the new model contains (13) and the introduction of L_R as a variable. The essential factor of rent seeking is that L_R becomes positive.

Let us start with a solution for an import restriction without rent seeking and ask what happens to the values of the variables when rent seeking is introduced. By assumption $M=\bar{M}$ is unchanged, so that L_D is unchanged. Therefore, $dL_A = -dL_R$, because the labor that enters rent seeking can only come from agriculture. Substituting into the total differential of (1) and using (6),

$$dF = dA = -A'dL_R < 0 \tag{14}$$

Agricultural production and food consumption are reduced by the introduction of rent seeking. Since the import level remains unchanged, rent seeking entails a welfare loss beyond that for an import restriction without rent seeking. The concavity of the agricultural production function results in a food loss that is less than proportional to decrements in L_A. Differentiating (5) totally,

$$0 = M_1 dp_M + M_2 dA \tag{15}$$

where M_1 and M_2 are the partial derivatives of (5) with respect to p_M and A, respectively. Solving (15) for dp_M, and substituting from (4) and (14),

$$dp_D = dp_M = \frac{M_2}{M_1} A' dL_R < 0 \tag{16}$$

since $M_1 < 0$ and $M_2 > 0$. The domestic cost of imports will be lower under rent-seeking competition. This follows from the decrease in the consumption of food relative to imports.

The results of (14) and (15) are not dependent upon the particular form of the equilibrium of the labor market. They hold for any specification of competitive rent seeking. Equation (13) serves to determine particular values for L_R and other variables of the system. The mere existence of competitive rent seeking is enough to determine the directions of change of the variables.

The above results are sufficient to indicate that, for any given level of import restrictions, competition among rent seekers is clearly inferior to the tariff equivalent of the restrictions, in that there could be more food consumed with no fewer imports under the latter case than the former. To the extent that rent seeking is competitive, the welfare cost of import restrictions is equal to the welfare cost of the tariff equivalent *plus the additional cost of rent-seeking activities*. Measurement of that excess cost is considered below.

The tariff-equivalent and rent-seeking equilibria are contrasted in Figure 3. Equilibrium under rent seeking will be at some point such as L, with the same consumption of imports, but smaller production and consumption of food than occurs under a tariff. The points K and C are the tariff-equivalent and free-trade equilibria, respectively. The line $D'D'$ cor-

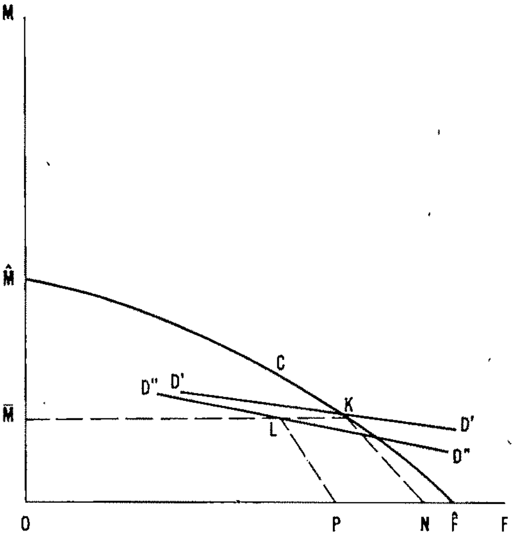


FIGURE 3. RENT-SEEKING IMPORT RESTRICTION

responds to the domestic price of imports in Figure 2, and the steeper line $D''D''$ corresponds to the lower domestic price of imports under competitive rent seeking.

So far, it has been shown that for any given level of import restriction, a tariff is Pareto-superior to competitive rent seeking, and the properties of rent-seeking equilibrium have been contrasted with those of the tariff-equivalent case in the absence of competition for the rents. A natural question is whether anything can be said about the properties of rent-seeking equilibrium in contrast to those of a free-trade equilibrium, which is, after all, the optimal solution. It has been seen that the number of persons engaged in distribution declines from free trade to import restriction without rent seeking, and increases as one goes from that situation to competition for import licenses. Likewise, agricultural output increases between free trade and the tariff-equivalent case, and declines between that and rent seeking. The question is whether any unambiguous signs can be placed on the direction of these changes between free trade and rent seeking and, in particular, is it possible that society might produce and consume less of both goods under rent seeking than under free trade?

The answer is that if inequality (12) is satisfied, the absolute number of persons (L_D+L_R) in distribution will increase going from a free-trade to a rent-seeking equilibrium. If import demand is more elastic, the number of persons in distribution will decline. Contrasted with a free-trade equilibrium, there would be less agricultural output *and* fewer imports when inequality (12) holds. If, with import restriction, the income from distribution $p_D^*\overline{M}$ is greater than distributors' income at free trade, more persons will be employed in distribution-cum-rent seeking with import restriction than are employed under free trade.

E. *Measuring the Welfare Loss from Rent Seeking*

A tariff has both production and consumption costs, and it has already been shown that rent seeking entails costs in addition to those of a tariff. Many forms of competition for rents, however, are by their nature difficult to observe and quantify and one might therefore question the empirical content of the result so far obtained.

Fortunately, there is a way to estimate the production cost of rent seeking. That cost, in fact, is equal to the value of the rents. This can be shown as follows. The rent per import license, r, is:

$$r = p_D - kA' \tag{17}$$

This follows because the labor required to distribute one unit of imports is k, which could be used in agriculture with a return A'. Note that at free trade r equals zero. A distributor could efficiently distribute an import and earn his opportunity cost in agriculture with zero rent. The total value of rents, R, with competitive rent seeking is thus the rent per unit of imports times the amount imported.

$$R = r\overline{M} = (p_D - kA')\overline{M} \tag{18}$$

Using (3) and (13),

$$\begin{aligned} R &= \left(p_D - \frac{kp_D\overline{M}}{L_D + L_R}\right)\overline{M} \\ &= p_D\left(1 - \frac{L_D}{L_D + L_R}\right)\overline{M} \\ &= \frac{p_D\overline{M}L_R}{L_D + L_R} \end{aligned} \tag{19}$$

Thus the total value of rents reflects the agricultural wage (A') times the number of rent seekers.

The value of rents reflects the value (at current prices) of the domestic factors of production which could be extracted from the economy with no change in the final goods and services available for society's utilization. Thus, if the value of rents is known, it indicates the volume of resources that could be transferred out of distribution and into other activities, with no loss of distributive services from an initial position of rent-seeking activity. The estimates of rents in India and Turkey, therefore, may be interpreted as the deadweight loss from quantitative restrictions in addition to the welfare cost of their associated tariff equivalents if one believes that there is competition for the rents.

The value of the rents overstates the increase in food output and consumption that could be attained with a tariff to the extent that the marginal product of labor in agriculture is diminishing, since the equilibrium wage will rise between the tariff and the competitive rent-seeking situation. In the case of a constant marginal product of labor in alternative uses, the value of rents will exactly measure foregone output.

F. *The Implications of Rent Seeking for Trade Theory*

Recognition of the fact of rent seeking alters a variety of conclusions normally obtained in the trade literature and examination of such cases is well beyond the scope of this paper. A few immediately derivable results are worth brief mention, however.

First, an import prohibition might be preferable to a nonprohibitive quota if there is competition for licenses under the quota. This follows immediately from the fact that a prohibition would release resources from rent seeking and the excess cost of domestic production might be less than the value of the rents. Second, one could not, in general, rank the tariff-equivalents of two (or more) quotas, since the value of rents is a function of both the amount of rent per unit (the tariff equivalent) and the volume of imports of each item.[12] Third, it has generally been accepted that the more inelastic domestic demand the less is likely to be the welfare cost of a given tariff. For the quota-cum-rents case, the opposite is true: the more price inelastic is demand. the greater will be the value of rents and the greater, therefore, the deadweight loss associated with rent seeking. Fourth, it is usually believed that competition among importers will result in a better allocation of resources than will a monopoly. If rent seeking is a possibility, however, creating a monopoly position for one importer will generally result in a higher real income if not in a preferable income distribution for society. Finally, devaluation under quantitative restrictions may have important allocation effects because it diminishes the value of import licenses, and hence the amount of rent-seeking activity, in addition to its effects upon exports.

III. Conclusions and Implications

In this paper, focus has been on the effects of competition for import licenses under a quantitative restriction of imports. Empirical evidence suggests that the value of rents associated with import licenses can be relatively large, and it has been shown that the welfare cost of quantitative restrictions equals that of their tariff equivalents plus the value of the rents.

While import licenses constitute a large and visible rent resulting from government intervention, the phenomenon of rent seeking is far more general. Fair trade laws result in firms of less-than-optimal size. Minimum wage legislation generates equilibrium levels of unemployment above the optimum with associated deadweight losses, as shown by John Harris and

---

[12] I am indebted to Bhagwati for pointing out this implication.

Michael Todaro, and Todaro. Ceilings on interest rates and consequent credit rationing lead to competition for loans and deposits and/or high-cost banking operations. Regulating taxi fares affects the average waiting time for a taxi and the percent of time taxis are idle, but probably not their owners' incomes, unless taxis are also licensed. Capital gains tax treatment results in overbuilding of apartments and uneconomic oil exploration. And so on.

Each of these and other interventions lead people to compete for the rents although the competitors often do not perceive themselves as such. In each case there is a deadweight loss associated with that competition over and above the traditional triangle. In general, prevention of that loss can be achieved only by restricting entry into the activity for which a rent has been created.

That, in turn, has political implications. First, even if they *can* limit competition for the rents, governments which consider they must impose restrictions are caught on the horns of a dilemma: if they do restrict entry, they are clearly "showing favoritism" to one group in society and are choosing an unequal distribution of income. If, instead, competition for the rents is allowed (or cannot be prevented), income distribution may be less unequal and certainly there will be less appearance of favoring special groups, although the economic costs associated with quantitative restrictions will be higher.

Second, the existence of rent seeking surely affects people's perception of the economic system. If income distribution is viewed as the outcome of a lottery where wealthy individuals are successful (or lucky) rent seekers, whereas the poor are those precluded from or unsuccessful in rent seeking, the market mechanism is bound to be suspect. In the United States, rightly or wrongly, societal consensus has been that high incomes reflect—at least to some degree—high social product. As such, the high American per capita income is seen as a result of a relatively free market mechanism and an unequal distribution is tolerated as a by-product. If, instead, it is believed that few businesses would survive without exerting "influence," even if only to bribe government officials to do what they ought in any event to do, it is difficult to associate pecuniary rewards with social product. The perception of the price system as a mechanism rewarding the rich and well-connected may also be important in influencing political decisions about economic policy. If the market mechanism is suspect, the inevitable temptation is to resort to greater and greater intervention, thereby increasing the amount of economic activity devoted to rent seeking. As such, a political "vicious circle" may develop. People perceive that the market mechanism does not function in a way compatible with socially approved goals because of competitive rent seeking. A political consensus therefore emerges to intervene further in the market, rent seeking increases, and further intervention results. While it is beyond the competence of an economist to evaluate the political impact of rent seeking, the suspicion of the market mechanism so frequently voiced in some developing countries may result from it.

Finally, all market economies have some rent-generating restrictions. One can conceive of a continuum between a system of no restrictions and a perfectly restricted system. With no restrictions, entrepreneurs would seek to achieve windfall gains by adopting new technology, anticipating market shifts correctly, and so on. With perfect restrictions, regulations would be so all-pervasive that rent seeking would be the only route to gain. In such a system, entrepreneurs would devote all their time and resources to capturing windfall rents.

While neither of these extreme types could ever exist, one can perhaps ask whether there might be some point along the continuum beyond which the market fails to perform its allocative function to any satisfactory degree. It will remain for further work to formalize these conjectures and to test their significance. It is hoped, however, that enough has been said to stimulate interest and research on the subject.

REFERENCES

J. **Bhagwati**, "On the Equivalence of Tariffs and Quotas," in his *Trade, Tariffs and Growth*, London 1969.

——— and P. Desai, *Planning for Industrialization: A Study of India's Trade and Industrial Policies Since 1950*, Cambridge 1970.

——— and A. Krueger, *Foreign Trade Regimes and Economic Development: Experience and Analysis*, New York forthcoming.

J. R. Harris and M. P. Todaro, "Migration, Unemployment, and Development: A Two-Sector Analysis," *Amer Econ. Rev.*, Mar. 1970, *60*, 126–42.

H. G. Johnson, "The Cost of Protection and the Scientific Tariff," *J. Polit. Econ.*, Aug. 1960, *68*, 327–45.

A. Krueger, *Foreign Trade Regimes and Economic Development: Turkey*, New York 1974.

J. B. Monteiro, *Corruption*, Bombay 1966.

G. Myrdal, *Asian Drama*, Vol. III, New York 1968.

M. P. Todaro, "A Model of Labor Migration and Urban Employment in Less Developed Countries," *Amer. Econ. Rev.*, Mar. 1969, *59*, 138–48.

R. Wraith and E. Simpkins, *Corruption in Developing Countries*, London 1963.

Government of India, Planning Commission, *Third Five Year Plan*, New Delhi, Aug. 1961.

Reserve Bank of India, *Report on Currency and Finance*, 1967–68.

Santhanam Committee, *Report on the Committee on Prevention of Corruption*, Government of India, Ministry of Home Affairs, New Delhi 1964.

The Empirical Relationships Between the Dividend and Investment Decisions of Firms

By EUGENE F. FAMA*

Franco Modigliani and Merton Miller (M-M) establish that in a perfect capital market[1] optimal investment decisions by a firm are independent of how such decisions are financed. This theorem has an important corollary: Investment decisions should never be determined by dividend decisions, and dividend decisions period-by-period need not be affected by investment decisions. The main goal of the present paper is to examine empirically the extent to which the dividend and investment decisions of individual firms are interrelated.[2]

Many models of investment and financing decisions based on the assumption of an imperfect capital market are available. Only the model of Phoebus Dhrymes and Mordecai Kurz is discussed here, since they provide what appears to be corroborating empirical evidence. Following John Meyer and Edmund Kuh, Dhrymes and Kurz propose a world in which, because of capital market imperfections, internal funds are a cheaper source of financing for the firm than new security issues, and dividends and investments are competing uses for limited internal funds. Moreover they hypothesize that firms not only allow investment decisions to affect dividend decisions, but that the desire to pay "reasonable" dividends causes investment decisions to be affected by dividend decisions. The Dhrymes-Kurz empirical evidence is discussed later.

I. Background

There is a substantial empirical literature on dividend and investment models. For dividends, the "partial adjustment" model of John Lintner has stood up well. In this model, the change, $\Delta D_{it}=D_{it}-D_{i,t-1}$, in the dividend of firm i from year $t-1$ to year t is given by

$$(1) \qquad \Delta D_{it} = \gamma_{1i}(D^*_{it} - D_{i,t-1}) + \epsilon_{it}$$

where ϵ_{it} is a random disturbance, γ_{1i} is the "speed-of-adjustment coefficient," and the target dividend D^*_{it} is assumed to be proportional to earnings P_{it}:

$$(2) \qquad D^*_{it} = \gamma_{2i}P_{it}$$

Combining equations (1) and (2) yields

$$(3) \qquad \Delta D_{it} = \beta_{1i}D_{i,t-1} + \beta_{2i}P_{it} + \epsilon_{it}$$

$$(4) \qquad \beta_{1i} = -\gamma_{1i}$$

$$(5) \qquad \beta_{2i} = \gamma_{1i}\gamma_{2i}$$

The Lintner model has fared well relative to its competitors in tests on aggregate data by John Brittain and in tests on data for individual firms by the author and Harvey Babiak.

The investment models most common

---

* Professor of finance, Graduate School of Business, University of Chicago. This project was financed by a grant from the National Science Foundation. I am grateful to David Booth and Stanley Smith for computational assistance, and to Calvin Siebert, George Borts, Nicholas Gonedes, Arthur Laffer, and Arnold Zellner for helpful comments.

[1] That is, a market where there are no transactions costs or taxes, all available information is costlessly available to everyone, and investors and firms are price takers.

[2] For a detailed discussion of the M-M model, see the book by Miller and the author, ch. 4.

in empirical work are close in spirit to the dividend model of equation (1). In the "flexible accelerator" model of Hollis Chenery and L. M. Koyck, the change in capital stock, $\Delta K_{it} = K_{it} - K_{i,t-1}$, is given by

$$\Delta K_{it} = \phi_{1i}(K^*_{it} - K_{i,t-1}) + \eta_{it} \tag{6}$$

In the Chenery-Koyck model the target capital stock K^*_{it} is proportional to output Q_{it}:

$$K^*_{it} = \phi_{2i}Q_{it} \tag{7}$$

Combining equations (6) and (7) yields,

$$\Delta K_{it} = \alpha_{1i}K_{i,t-1} + \alpha_{2i}Q_{it} + \eta_{it} \tag{8}$$

$$\alpha_{1i} = -\phi_{1i} \tag{9}$$

$$\alpha_{2i} = \phi_{1i}\phi_{2i} \tag{10}$$

There is, however, disagreement in the literature concerning whether output Q_{it} is the appropriate variable in (7). For example, Jan Tinbergen and others favor profits P_{it}. But models using output, or some variant of output, have performed well relative to these competitors.[3]

Many types of relationships between dividend and investment decisions are consistent with the Modigliani-Miller theorem. In fact, the theorem only rules out cases where firms allow dividend decisions to affect investment decisions. In a perfect capital market, investment decisions are separable from financing decisions. Firms are free to choose dividend decisions that mirror or do not mirror the details of investment decisions to whatever extent they like, but optimal investment decisions are not determined by dividend decisions.

At one extreme, there might be no detectable relationships between the two decisions. Or, firms might choose dividend policies that are on average complementary to optimal investment decisions, but make no effort to adjust dividends period-by-period to reflect the deviations of investments from "normal." For example, firms with high target ratios of capital stock to output may choose low target ratios of dividends to profits. Or, firms that adjust quickly to discrepancies between target and actual capital stock may choose slower speeds of adjustment in their dividend decisions. But firms may not make their dividend decisions conform period-by-period to their investment decisions. Specifically, the disturbances of the dividend and investment equations might be independent. Finally, it is even consistent with the M-M theorem that the disturbances of dividend and investment equations are dependent, as long as the direction of causation is from investment to dividends.

These and other types of relationships between dividend and investment decisions are examined in this paper. The approach is as follows. Many variants of equations (1) and (6) are proposed and tested in the literature. The first step is to screen various models, primarily on the basis of the quality of the predictions they provide. Considerable effort is spent on this stage, and results on the empirical properties of dividend and investment models, which seem interesting, are obtained. Models chosen from this screening are then used to present the tests for relationships between dividend and investment decisions. First, simultaneous equations models of dividend and investment decisions are studied. Then we look for relationships between the coefficient estimates and between the residuals of dividend and investment regressions. We cannot reject the hypothesis that, in general, there is a rather complete independence between the dividend and investment decisions of individual firms.

---

[3] For a more extensive review of the empirical investment literature, see Dale Jorgenson (1971).

II. Regressions and Predictions for Single Equation Models

The data file for this study is annual financial statement information on 900 major industrial firms for the period 1946–68, as reported on the Compustat tapes of the Standard Statistics Corporation. But the actual sample used includes only the 298 firms for which complete data on all variables needed in the various tests are available for the entire 23-year period.

The estimated dividend and investment regressions are of the general form,[4]

$$(11)\quad \Delta D_{it} = \hat{\beta}_{1i}D_{i,t-1} + \hat{\beta}_{2i}P_{it} + \hat{\beta}_{3i}P_{i,t-1} + \hat{\beta}_{4i}G_{it} + \hat{\beta}_{5i}N_t + \hat{\varepsilon}_{it}$$

$$(12)\quad \Delta K_{it} = \hat{\alpha}_{1i}K_{i,t-1} + \hat{\alpha}_{2i}Q_{it}(\text{or } \hat{\alpha}_{2i}P_{it}) + \hat{\alpha}_{3i}G_{it} + \hat{\alpha}_{4i}G_{i,t-1} + \hat{\alpha}_{5i}N_t + \hat{\eta}_{it}$$

$$(13)\quad \Delta K_{it} = \hat{\alpha}'_{1i}\Delta K_{i,t-1} + \hat{\alpha}_{2i}\Delta Q_{it} + \hat{\alpha}_{3i}\Delta G_{it} + \hat{\alpha}_{4i}\Delta N_t + \hat{\xi}_{it}$$

where the subscripts refer to firm i and year t, and the variables are measured as

K_{it} = Net plant and equipment
Q_{it} = Sales plus change in inventories
P_{it} = "Available for common" = Net income − Preferred dividends
G_{it} = Depreciation
N_t = Gross national product x 10^{-10}

All firm-related variables are measured on a per share basis, but with the appropriate adjustments to abstract from the effects of "changes" in the number of shares outstanding caused by stock splits and stock dividends.[5]

Equation (13) follows the suggestion of Dale Jorgenson and Calvin Siebert who recommend fitting investment models in differenced form.[6] Including the lagged profit term in the dividend equation is suggested by the results reported by Babiak and the author. Including depreciation in the equations is in line with the "cash flow" dividend model of Brittain and the "internal funds" investment model of Dhyrmes and Kurz, and others. Finally, following John Gould and Roger Waud, Gross National Product (*GNP*) is included in the regressions as a shift variable.

A. *Prediction Results*

The initial task is to judge the power of various versions of the dividend and investment regressions (11)–(13) by comparing cross-sectional distributions of prediction errors. Coefficient estimates for each regression equation are obtained for each of the 298 firms, using the first 21 years (1946–66) of data. These coefficient estimates are then used to make conditional predictions of changes in dividends and capital stock for each firm for the years 1967 and 1968. The predictions are conditional in that they use values of the explanatory variables appropriate for 1967 and 1968. Prediction errors are computed as differences between actual changes in dividends (or capital stock) from 1966–67 (or from 1967–68) and the changes predicted by the regression models. Since the dispersion of the distribution of the prediction error for any given model is likely to vary from firm to firm, in constructing cross-sectional distributions, the

---

[4] There are, of course, many statistical problems associated with regression models where lagged values of the dependent variable appear as explanatory variables. In the article by Babiak and the author, these problems are discussed in detail and studied via simulations on models much like (11)–(13).

[5] There are pros and cons for the use of per share data. But many of the tests were also done without adjusting the variables to a per share basis, and the results were never very different.

[6] The differenced form is obtained by adding $K_{i,t-1}$ to both sides of an investment equation, and then differencing. For example, applying this procedure to (8) yields

$$\Delta K_{it} = \alpha'_{1i}\Delta K_{i,t-1} + \alpha_{2i}\Delta Q_{it} + \xi_{it}$$

$$\alpha'_{1i} = 1 + \alpha_{1i} = 1 - \phi_{1i}$$

$$\xi_{it} = \eta_{it} - \eta_{i,t-1}$$

prediction errors are measured in units of the standard deviation of the dependent variable. Thus the dividend prediction errors are standardized as follows:

$$\hat{u}_{i,67} = \frac{\hat{e}_{i,67}}{s(\Delta D_i)}$$

where $\hat{e}_{i,67}$ is the prediction error for firm i for 1967 for a particular version of the dividend model, and where $s(\Delta D_i)$ is the standard deviation of the annual changes in the dividend of firm i over the period 1946–66. The investment prediction errors are likewise standardized by $s(\Delta K_i)$.

The results of the prediction tests are summarized in Table 1. Each regression model is first given a number and then the explanatory variables are shown. For each model and for each of the prediction years 1967 and 1968, the remaining columns of the table then show the average prediction error (for example, $\sum_i \hat{u}_i/298$), the average absolute value of the prediction errors (for example, $\sum_i |\hat{u}_i|/298$), the mean square prediction error (for example, $\sum_i \hat{u}_i^2/298$), and the standard deviation of the prediction errors. Results are also shown for a naive model (Average Change) in which the prediction is that the changes in dividends (or capital stock) for 1967 and 1968 are equal to the average annual change for the period 1946–66.[7,8]

Dividends

For both the 1967 and 1968 predictions, and on every measure of dispersion of pre-

TABLE 1—SUMMARY OF CROSS-SECTIONAL DISTRIBUTIONS OF PREDICTION ERRORS FOR 1967 AND 1968

		Predictions for 1967				Predictions for 1968			
			Dispersion Measures				Dispersion Measures		
Model	Explanatory Variables[a]	Average	Average Absolute	Mean Square	Standard Deviation	Averate	Average Absolute	Mean Square	Standard Deviation
Dividends									
D1	D_{t-1}, P_t	−.00	.83	1.24	1.11	−.23	.81	1.16	1.05
D2	D_{t-1}, P_t, P_{t-1}	−.10	.88	1.41	1.18	−.29	.87	1.39	1.14
D3	$D_{t-1}, P_t, P_{t-1}, G_t$	−.10	.91	1.53	1.23	−.28	.98	1.83	1.32
D4	$D_{t-1}, P_t, P_{t-1}, G_t, N_t$	−.09	.94	1.57	1.25	−.21	1.16	2.60	1.60
	Average Change	.37	.90	1.69	1.24	.16	.78	1.38	1.16
Investment									
K1	K_{t-1}, Q_t	.89	1.74	6.84	2.46	.87	2.06	11.25	3.24
K2	K_{t-1}, Q_t, G_t	.85	1.77	7.33	2.57	.74	2.13	10.70	3.18
K3	$K_{t-1}, Q_t, G_t, G_{t-1}$	.69	1.71	6.92	2.54	.68	2.21	12.95	3.53
K4	$K_{t-1}, Q_t, G_t, G_{t-1}, N_t$	.62	1.73	7.09	2.59	.57	2.41	15.48	3.89
K5	$\Delta K_{t-1}, \Delta Q_t$	.86	1.72	7.12	2.53	.49	1.84	9.31	3.01
K6	$\Delta K_{t-1}, \Delta Q_t, \Delta G_t$	.78	1.60	6.32	2.39	.47	1.87	9.88	3.11
K7	$\Delta K_{t-1}, \Delta Q_t, \Delta G_t, \Delta N_t$	.41	1.54	5.79	2.37	.43	1.91	13.16	3.60
K8	K_{t-1}, P_t	.93	1.74	7.13	2.50	.93	1.97	10.79	3.15
	Average Change	1.47	1.88	8.37	2.49	1.44	2.05	11.64	3.09

[a] To conserve space, the subscript i, which should appear on all explanatory variables except N_t and ΔN_t, is omitted.

---

[7] The models reported in Table 1 were also fit with constants. The predictive power of such models is poor. Dividend models with constants have mean square errors in 1967 that are 5–10 percent greater than those for corresponding models without constants, and 20–30 percent greater in 1968. For investment models, the margins of superiority of models with constants suppressed is 5–10 percent in 1967 and 10–30 percent in 1968.

[8] Cross-sectional distributions of prediction errors were also constructed for dividend and investment models in which the prediction error for a firm is taken from the regression model in Table 1 with the highest adjusted coefficient of determination for that firm. In prediction tests, these highest R^2 models do poorly relative to the better predictive models in Table 1.

diction errors, the two-variable Lintner model (**D1**) does better than all other dividend regression models. Moreover, the dividend predictions deteriorate systematically as additional variables are included in the model. Thus, depreciation, lagged profits and *GNP* do not in general help to explain the dividend decisions of individual firms. Models that include these variables do not even do well in comparison with the naive Average Change model.

Investments

Unlike the dividend models, the rankings of the investment models show little consistency across the two prediction years. In five of six comparisons, the differenced models **K5**, **K6**, and **K7** have lower mean square prediction errors than the corresponding models, **K1**, **K2**, and **K4**, in which the explanatory variables appear as levels. But the four-variable differenced model **K7**, which does best in the predictions for 1967, is among the worst in the predictions for 1968.

As in the dividend models, *GNP* and depreciation do not seem to be systematically important in explaining changes in capital stock. In both prediction years, model **K3**, which does not include *GNP*, has lower mean square error than model **K4**, which does, while model **K7** does better than **K6** in 1967 but worse in 1968. As for depreciation, in five of six comparisons the two-variable output model **K1** has lower mean square error than the models **K2**–**K4** which include depreciation; and the differenced form of the two-variable output model, **K5**, wins two of four comparisons with models **K6** and **K7**. As in the case of dividends, investment models that include depreciation and *GNP* as explanatory variables generally do not even do well relative to the naive Average Change model.[9]

Finally, the three measures of cross-sectional dispersion of prediction errors do not allow a clear choice between the two-variable output investment model, **K1**, and the profits model **K8**: The output model is slightly better in the 1967 predictions, while the profits model is slightly better in 1968. In terms of consistency of predictive performance for the two years, the two-variable output model, **K5**, which is just the differenced form of **K1**, seems slightly better than both **K1** and **K8**.[10]

Other Prediction Issues

Several features of the results in Table 1 deserve further attention. For example, the predictions of the investment regression models seem "biased"; in both prediction years, on average, actual ΔK_{it} exceeds $\widehat{\Delta K}_{it}$ by more than $.5s(\Delta K_i)$. Also, given that the underlying standardized prediction errors are in units of $s(\Delta K_i)$, the measures of the dispersion of the investment prediction errors seem large. Table 2 is meant to shed light on these and other issues.

Columns (2) and (3) of Table 2 show for each year t the average across firms of $[(\Delta D_{it}-\overline{\Delta D}_i)/s(\Delta D_i)]^2$ and $[(\Delta K_{it}-\overline{\Delta K}_i)/s(\Delta K_i)]^2$, where $\overline{\Delta D}_i$, $\overline{\Delta K}_i$, $s(\Delta D_i)$, and $s(\Delta K_i)$ are the means and standard deviations of ΔD_{it} and ΔK_{it} over the period 1946–66. It is apparent from column (3) of Table 2 that the cross-sectional dispersion of ΔK_{it} has increased through time—most noticeably in the last several years of the period. The high dispersion of the prediction errors for the investment models in Table 1 is a direct consequence of this fact. On the other hand, the cross-sectional dispersion of dividend changes does not seem

---

[9] A measure of the average change in capital stock for the 298 firms in the sample was also tried as a shift variable, with results no better than those for *GNP*.

[10] Many other profits models, including differenced models analogous to **K5**–**K7**, were also tried, with poor results. Differenced forms of the dividend models also performed poorly.

TABLE 2—SOME STATISTICS ON THE BEHAVIOR OF DIVIDENDS AND INVESTMENTS THROUGH TIME

	Cross-Sectional Averages of Squared Deviations from Time-Series Means		Cross-Sectional Averages of Deviations from Time-Series Means	
(1) Year (t)	(2) ΔD_{it}	(3) ΔK_{it}	(4) ΔD_{it}	(5) ΔK_{it}
1947	1.36	.51	.22	.08
1948	1.14	.67	.13	.11
1949	1.10	.42	−.31	−.25
1950	1.54	.48	.45	−.33
1951	1.26	.55	−.20	.07
1952	.85	.58	−.49	−.12
1953	.65	.69	−.38	−.18
1954	.75	.62	−.09	−.30
1955	.88	.75	.17	−.27
1956	.75	.95	.20	.23
1957	.78	.79	−.11	.15
1958	1.19	.81	−.60	−.24
1959	.76	.95	−.08	−.29
1960	.68	.73	.01	−.06
1961	.63	.92	−.31	−.27
1962	.51	.72	−.15	−.20
1963	.56	1.06	−.05	−.16
1964	.83	1.23	.31	.14
1965	1.19	1.90	.59	.63
1966	1.57	3.67	.63	1.27
1967	1.69	8.37	.37	1.47
1968	1.38	11.64	.16	1.44

to have changed much, the average values of $[(\Delta D_{it}-\overline{\Delta D}_i)/s(\Delta D_i)]^2$ are much the same in the first and last few years of the period.

Columns (4) and (5) of Table 2 show for each year t the average across firms of $(\Delta D_{it}-\overline{\Delta D}_i)/s(\Delta D_i)$ and $(\Delta K_{it}-\overline{\Delta K}_i)/s(\Delta K_i)$. The numbers are systematically positive from 1964 onward, indicating a "boom" in both dividends and investment during these years, with the boom in investment being stronger. Since the average prediction errors for the dividend regression models in Table 1 are close to zero, the boom in dividends just reflects a normal adjustment to the behavior of the explanatory variables in the regression models in those years (i.e., a boom in profits). But the boom in investment goes beyond the normal adjustment to values of the explanatory variables (for example, output). The average prediction errors for 1967 and 1968 for the investment regression models in Table 1 are somewhat smaller than the last two numbers in column (5) of Table 2, but the average prediction errors for the regression models are nevertheless substantially positive in both years. This suggests that perhaps the regression models overlook variables that are important in determining investment expenditures, at least in the two prediction years 1967 and 1968. Or, perhaps there are upward shifts through time in target capital to output ratios as a consequence of increasingly capital-intensive technologies for the firms in the sample.

B. *Details on the Regressions*

Table 3 shows means, standard deviations, and five fractiles of the cross-sectional distributions of the firm-by-firm coefficient estimates, *t*-statistics for coefficient estimates, coefficients of determination R_i^2 adjusted for degrees of freedom, and first-order serial correlations $\hat{\rho}_i$ and studentized ranges SR_i of residuals for the two-variable Lintner dividend model **D1** and for the two-variable investment models **K1**, **K5**, and **K8**. Also summarized are the cross-sectional distributions of firm-by-firm estimates of speed-of-adjustment coefficients (SOA_i), the target payout ratio (TPO_i) for the dividend model **D1**, the target capital-to-output ratios (TKQ_i) for the investment models **K1** and **K5**, and the target capital-to-profit ratio (TKP_i) for the investment model **K8**. The estimates of speed-of-adjustment coefficients and of the various target ratios are obtained from equations (1)–(13), with coefficient estimates replacing population values.

Table 3 gives additional insight into how well the regression models represent dividend and investment decisions. For exam-

TABLE 3—DETAILED RESULTS FOR SELECTED DIVIDEND AND INVESTMENT MODELS

Panel A (Model D1)

$$\Delta D_{it} = \hat{\beta}_{1i} D_{i,t-1} + \hat{\beta}_{2i} P_{it} + \hat{\epsilon}_{it}$$

	$\hat{\beta}_{1i}$	$\hat{\beta}_{2i}$	R_i^2	$t(\hat{\beta}_{1i})$	$t(\hat{\beta}_{2i})$	$\hat{\rho}_i$	SR_i	SOA_i	TPO_i
Mean	−.269	.154	.42	−3.30	4.15	−.053	4.17	.269	.591
Standard Deviation	.149	.087	.22	1.78	1.98	.251	.54	.149	.178
5 Fractiles									
.10	−.470	.060	.11	−5.29	1.94	−.378	3.50	.104	.401
.30	−.339	.108	.32	−3.91	3.09	−.186	3.84	.182	.525
.50	−.251	.141	.42	−3.09	3.82	−.054	4.16	.251	.584
.70	−.182	.181	.54	−2.34	4.84	.072	4.46	.339	.660
.90	−.104	.265	.72	−1.39	6.85	.295	4.88	.470	.779

Panel B (Model K1)

$$\Delta K_{it} = \hat{\alpha}_{1i} K_{i,t-1} + \hat{\alpha}_{2i} Q_{it} + \hat{\eta}_{it}$$

	$\hat{\alpha}_{1i}$	$\hat{\alpha}_{2i}$	R_i^2	$t(\hat{\alpha}_{1i})$	$t(\hat{\alpha}_{2i})$	$\hat{\rho}_i$	SR_i	SOA_i	TKQ_i
Mean	−.187	.089	.25	−1.73	2.55	.156	4.09	.187	−.424
Standard Deviation	.249	.134	.23	2.14	2.19	.235	.49	.249	11.941
5 Fractiles									
.10	−.451	.005	−.00	−3.50	.32	−.122	3.47	−.053	.029
.30	−.270	.030	.12	−2.58	1.76	.015	3.78	.092	.180
.50	−.182	.053	.23	−1.82	2.46	.152	4.07	.182	.275
.70	−.092	.091	.34	−1.01	3.18	.262	4.40	.270	.496
.90	.053	.219	.59	.65	4.37	.492	4.72	.451	1.056

Panel C (Model K5)

$$\Delta K_{it} = \hat{\alpha}_{1i} \Delta K_{i,t-1} + \hat{\alpha}_{2i} \Delta Q_{it} + \hat{\xi}_{it}$$

	$\hat{\alpha}_{1i}$	$\hat{\alpha}_{2i}$	R_i^2	$t(\hat{\alpha}_{1i})$	$t(\hat{\alpha}_{2i})$	$\hat{\rho}_i$	SR_i	SOA_i	TKQ_i
Mean	.400	.142	.14	2.04	2.37	−.163	4.06	.600	.253
Standard Deviation	.314	.215	.32	1.70	2.29	.188	.46	.314	.471
5 Fractiles									
.10	.020	.012	−.23	.08	.29	−.422	3.45	.207	.006
.30	.237	.047	−.03	1.08	1.28	−.251	3.78	.432	.081
.50	.385	.074	.12	1.78	2.08	−.153	4.05	.619	.140
.70	.568	.143	.31	2.70	3.07	−.068	4.32	.770	.261
.90	.798	.365	.58	4.40	4.45	.086	4.72	.980	.731

Panel D (Model K8)

$$\Delta K_{it} = \hat{\alpha}_{1i} K_{i,t-1} + \hat{\alpha}_{2i} P_{it} + \hat{\eta}_{it}$$

	$\hat{\alpha}_{1i}$	$\hat{\alpha}_{2i}$	R_i^2	$t(\hat{\alpha}_{1i})$	$t(\hat{\alpha}_{2i})$	$\hat{\rho}_i$	SR_i	SOA_i	TKP_i
Mean	−.043	.664	.17	−.55	1.36	.077	4.18	.043	−1.372
Standard Deviation	.132	.792	.22	1.39	1.40	.228	.51	.132	118.656
5 Fractiles									
.10	−.174	.044	−.06	−2.05	.18	−.214	3.53	−.067	−14.152
.30	−.083	.317	.02	−1.22	1.18	−.043	3.84	−.003	1.787
.50	−.036	.489	.14	−.68	1.87	.075	4.17	.036	6.198
.70	.003	.768	.27	.05	2.46	.198	4.51	.083	12.502
.90	.067	1.484	.49	1.12	3.53	.388	4.85	.174	28.059

ple, the cross-sectional distributions of studentized ranges SR_i do not indicate that severe departures from normality are a general characteristic of either the dividend or the investment models. In samples of size 19 from a normal population, the .10 fractile of the distribution of the studentized range is 3.29, while the .90 fractile is 4.32. In the cross-sectional distributions of SR_i in Table 3, the .10 and .90 fractiles are slightly to the right of these values, indicating that the distributions of residuals are slightly "thick-tailed" relative to the normal.

Most of the results in Table 3 suggest that the Lintner dividend model stands up somewhat better to the data than the investment models. Thus, the cross-sectional distribution of R_i^2 for the dividend model is substantially to the right of those for the investment models. Moreover, in the dividend model, the two cross-sectional distributions of t-statistics indicate that coefficient estimates are generally large relative to their standard errors. For the investment models, this is not so generally the case.

Serial correlation of residuals is also more serious in the investment models. In the dividend model the cross-sectional distribution of $\hat{\rho}_i$ is consistent with what Babiak and the author observe in simulations of similar models when the true serial correlation is zero, and inconsistent with what we observe when the true serial correlation deviates from zero by as little as ±.2.[11] In the investment regressions, the cross-sectional distributions of $\hat{\rho}_i$ for the profits model **K8** is quite consistent with true first-order serial correlation of about .2, while .2 would be low as an estimate of the true serial correlation for the output model **K1**. On the other hand, the cross-sectional distributions of $\hat{\rho}_i$ for model **K5** (which is just the differenced form of **K1**) is consistent with true serial correlation of about −.2. Positive serial correlation of residuals implies that the usual estimates of residual variances are downward biased estimates of disturbance variances, while the reverse holds in the case of negative serial correlation. This would explain why the investment model **K1** yields slightly higher values of R_i^2 than **K5**, but **K5** is as good or better in predictions.

TABLE 4—CROSS-SECTIONAL DISTRIBUTIONS OF AVERAGE DIVIDEND PAYOUT RATIOS (APO_i), AVERAGE CAPITAL-TO-OUTPUT RATIOS (AKQ_i), AND AVERAGE CAPITAL-TO-PROFIT RATIOS (AKP_i)

	APO_i	AKQ_i	AKP_i
Mean	.541	.646	6.967
Standard Deviation	.338	.672	14.681
5 Fractiles			
.10	.323	.165	2.036
.30	.447	.294	3.782
.50	.515	.404	5.401
.70	.578	.682	7.676
.90	.718	1.419	13.036

If the dividend and investment processes are to be interpreted in terms of the models of equations (1)–(10), there should be some association between target payout TPO_i, capital-to-output TKQ_i, and capital-to-profit TKP_i ratios estimated from regressions and the corresponding observed average payout APO_i, capital-to-output AKQ_i, and capital-to-profit AKP_i ratios. Table 4 summarizes the cross-sectional distributions of APO_i, AKQ_i, and AKP_i. The average payout ratio APO_i is just the average of the year-by-year values of the ratio of dividends to profits for firm i while AKQ_i and AKP_i are likewise averages of ratios of capital stock to output and to profits, respectively.

For the dividend model, the cross-sectional distribution of the target payouts in Table 3 is fairly comparable to that of the average payouts in Table 4. In the output investment models **K1** and **K5**, however, the target capital-to-output ratios are systematically low relative to the observed average capital-to-output ratios. Likewise, in the profits investment model **K8** there is little similarity between the cross-sectional distributions of target and average capital-to-profit ratios.

---

[11] In regression models where lagged values of the dependent variable appear as explanatory variable, sample serial correlations of residuals are biased toward zero when the true serial correlation is nonzero. Thus the reliance on the results of Babiak and the author. We used simulations to study the bias under different assumptions about the magnitude of the true serial correlation.

There is some evidence, however, that the differenced investment model **K5** is better on this score than models **K1** and **K8**. For model **K1** the correlation across the 298 firms of paired values of TKQ_i and AKQ_i is $-.003$ and for **K8** the corresponding correlation between paired values of TKP_i and AKP_i is $-.078$. For model **K5**, the correlation between TKQ_i and AKQ_i is .371, which compares favorably with the correlation coefficient of .211 between APO_i and TPO_i for the dividend model **D1**.[12]

A corollary of the Modigliani-Miller theorem on the independence of investment and financing decisions is that in a perfect capital market, investors are indifferent to a firm's dividend policy. Thus, beyond noting that the estimates are positive as predicted by the Lintner partial adjustment model, the "reasonableness" of the estimated speed-of-adjustment coefficients in the dividend regressions is impossible to judge. In the investment regressions we are in a slightly better position since we can at least make a rough comparison of the SOA_i estimates in Table 3 with the independent survey results of Thomas Mayer concerning the average time between initiation and completion of investment projects.

On this basis, the SOA_i obtained from the differenced output model **K5** are higher and as a result somewhat more reasonable than the estimates obtained from the output model **K1** and the profits model **K8**. For model **K1**, the average value of $SOA_i = -\hat{a}_{1i}$ is .187. In other words, this model estimates that for the average firm less than 20 percent of the adjustment to the difference between the desired and actual levels of capital stock occurs in any given year. Thus when a gap between desired and actual capital stock arises, many years must pass before this gap is substantially closed. And if the speed-of-adjustment estimates from the output model **K1** are considered too low, then the estimates obtained from the profits model **K8** (which average only .04) must be considered even more unreasonably low. But for the differenced model **K5** the average value of $SOA_i = 1 - \hat{a}_{1i}$ is .60. Thus for the average firm 60 percent of the adjustment to the difference between the desired and actual capital stock occurs in any given year, and when a gap between the desired and actual capital stock arises, it is substantially closed in two to three years. These numbers are more in line with the findings of Mayer.

C. *Evaluating the Regressions*

The results of this paper with respect to dividend models are consistent with those of Lintner, Brittain, and Babiak and the author. Specifically, the Lintner model consistently performs well relative to its competitors.

---

[12] Correlations of .371 and .211 may seem low, but keep in mind that the target ratios are computed as ratios of regression coefficients that contain nontrivial sampling errors. The downward bias of the estimates of target capital-to-output ratios for the investment model **K5**, however, is more difficult to explain, since the simulations of Babiak and the author on models much like those studied here indicate that "lagged variable bias" in the regression coefficients does not lead to bias in estimated target ratios—a finding that seems to be substantiated in the results for the dividend model **D1**. Perhaps the problem with the estimated target capital-to-output ratios is related to the previous suggestion, drawn from the prediction results in Tables 1 and 2, that true target capital-to-output ratios rose during the sample period as a consequence of increasingly capital-intensive technologies.

One might also argue that since there is considerable uncertainty about which model is best and even about whether the results for the "better" models can reasonably be interpreted in terms of the partial adjustment mechanisms and target capital ratios that form their theoretical foundations, the results of this paper with respect to investment models are also consistent with the previous literature in this area. Moreover, the investment models reported in Tables 1 and 3 are a small subset of those that were tried. For example, following the suggestion of Dale Jorgenson (1966) models including additional lags of explanatory variables were tested. Following the work of Robert Eisner, many "mixed" models that include measures of both output and profits as explanatory variables were tried. These models did worse, or at least no better, than those reported in Table 1.

Since we must make do with what we have, when we turn now to the tests of the relationships between dividend and investment decisions, attention is focused on the two-variable Lintner dividend model **D1** (a choice made with some confidence) and the two-variable output investment model **K1** and its differenced form **K5** (choices made with less confidence). But the results for these models are representative. All further tests were carried out for all the models in Table 1. For the conclusions to be drawn from these tests, the specific choice of models used in reporting makes no difference.

III. Simultaneous Equations Models

Applying two-stage least squares (*2SLS*) and three-stage least squares (*3SLS*) to cross-section data for individual firms, Dhrymes and Kurz find what seems to be strong evidence that investment decisions affect dividend decisions (which is consistent with the Modigliani-Miller perfect market model) and that dividend decisions affect investment decisions (which is inconsistent with the M-M world). They conclude that their evidence is consistent with an imperfect capital market in which internal funds are a cheaper source of financing for the firm than new security issues, and where investments and dividends are competing uses for limited internal funds. Moreover, their work seems to suggest that single equation models fit with ordinary least squares (*OLS*) are likely to give results much different from those obtained with simultaneous equations models fit by *2SLS* or *3SLS*. This section applies simultaneous equations models to our time-series data for individual firms. The findings are in complete opposition to those of Dhrymes and Kurz.

For dividends, the structural equation, corresponding to the Lintner model **D1**, is

$$\Delta D_{it} = \beta_{1i}D_{i,t-1} + \beta_{2i}P_{it} + \beta_{3i}\Delta K_{it} + \epsilon_{it} \tag{14}$$

while for investment, two structural equations, corresponding to model **K1** and its differenced form **K5**, are used:

$$\Delta K_{it} = \alpha_{1i}K_{i,t-1} + \alpha_{2i}Q_{it} + \alpha_{3i}D_{it} + \eta_{it} \tag{15}$$

$$\Delta K_{it} = \alpha'_{1i}\Delta K_{i,t-1} + \alpha_{2i}\Delta Q_{it} + \alpha_{3i}\Delta D_{it} + \xi_{it} \tag{16}$$

In fitting these equations by *2SLS*, the "instrumental variable" estimates of D_{it} and ΔD_{it} used in (15) and (16) are taken from the *OLS* regressions of the dividend model **D1** in Table 3 and are labelled D_{it}(**D1**) and $\widehat{\Delta D}_{it}$(**D1**). Instrumental variable estimates of ΔK_{it} in (14) are taken from models **K1** and **K5**, and are labelled $\widehat{\Delta K}_{it}$(**K1**) and $\widehat{\Delta K}_{it}$(**K5**). With these two instrumental variables, two separate sets of *2SLS* fits of (14) are obtained.

The simultaneous equations models are judged primarily by comparing the cross-

TABLE 5—SUMMARY OF CROSS-SECTIONAL DISTRIBUTIONS OF PREDICTION ERRORS: *2SLS* MODELS

			Predictions for 1967				Predictions for 1968			
				Dispersion Measures				Dispersion Measures		
Model	Explanatory Variables	*2SLS* Variables	Average	Average Absolute	Mean Square	Standard Deviation	Average	Average Absolute	Mean Square	Standard Deviation
Dividends										
D1	$D_{i,t-1}, P_{it}$	$\widehat{\Delta K_{it}}(K1)$	.05	.90	1.42	1.19	−.14	.95	1.66	1.28
D1	$D_{i,t-1}, P_{it}$	$\widehat{\Delta K_{it}}(\mathbf{K5})$	.08	.88	1.39	1.18	−.14	.98	1.99	1.40
Investment										
K1	$K_{i,t-1}, Q_{it}$	$D_{it}(\mathbf{D1})$	.82	1.76	6.73	2.46	.83	2.08	10.82	3.18
K5	$\Delta K_{i,t-1}, \Delta Q_{it}$	$\widehat{\Delta D_{it}}(\mathbf{D1})$	.83	1.74	7.42	2.59	.61	1.83	11.26	3.30

sectional distributions of prediction errors obtained from the *2SLS* regressions with those obtained for the corresponding single equation models fit by *OLS* that are reported in Table 1. Thus, as in the *OLS* regressions, the first twenty-one years of data (1946–66) are used in the *2SLS* regressions, with the data for 1967 and 1968 then used for predictions. The *2SLS* predictions are summarized in Table 5.

For both prediction years the single equation results in Table 1 for the dividend model **D1** are better on every measure of dispersion of prediction errors than the simultaneous equations results in Table 5.[13] The investment predictions likewise do not indicate that simultaneous equation models systematically improve on single equation results. The *OLS* estimates of **K5** seem to provide slightly better predictions than the *2SLS* estimates, while *OLS* and *2SLS* estimates for **K1** do about equally well.

For each of the models in Table 5, Table 6 shows averages across the 298 firms of the *2SLS* coefficient estimates, their t-statistics, and adjusted coefficients of determination. Although t-statistics and coefficients of determination are somewhat ambiguous in *2SLS* (and thus the emphasis on predictions), comparison of Table 6 with the *OLS* results for corresponding single equation models in Table 3 seems also to support the conclusion that the simultaneous equations models do not systematically improve on single equation models. For variables that the single and simultaneous equations models have in common, there is not much change in average values of estimated coefficients. More interesting, average values of R^2 also do not change much from Table 3 to Table 6, which simply reflects the fact that the instrumental variables in the *2SLS* regressions have little explanatory power. More specifically, in the *2SLS* dividend regressions the averages of the coefficients of $\widehat{\Delta K_{it}}(\mathbf{K1})$ and $\widehat{\Delta K_{it}}(\mathbf{K5})$ are negative—as would be predicted by the Dhrymes-Kurz hypothesis that dividends and investments are competing uses for limited internal funds—but average t-statistics for these coefficients are close to zero. More important, about 40 percent of the firm-by-firm values of the coefficients are positive and fewer than 5 percent have t-statistics less than -2.0. In the *2SLS* investment regressions about 60 per cent of the coefficients of $D_{it}(\mathbf{D1})$ and $\Delta D_{it}(\mathbf{D1})$

---

[13] To test whether gross investment $I_{it}=\Delta K_{it}+G_{it}$ affects dividend decisions, *2SLS* fits for dividend models that include both depreciation, G_{it}, and an instrumental variable estimate of ΔK_{it} as explanatory variables have been obtained. These models produced systematically poorer predictions than dividend models that do not include G_{it} as an explanatory variable—a result also previously observed in Table 1.

TABLE 6—SUMMARY OF *2SLS* REGRESSION RESULTS

Model	Dependent Variable	Explanatory Variables	Instrumental Variable	$\overline{\hat{\beta}_1}$	$\overline{\hat{\beta}_2}$	$\overline{\hat{\beta}_3}$	$\overline{R}^2$	$\bar{t}(\hat{\beta}_1)$	$\bar{t}(\hat{\beta}_2)$	$\bar{t}(\hat{\beta}_3)$
D1	ΔD_{it}	$D_{i,t-1}, P_{it}$	$\widehat{\Delta K_{it}}$(K1)	−.288	.167	−.011	.44	−3.16	3.30	−.19
D1	ΔD_{it}	$D_{i,t-1}, P_{it}$	$\widehat{\Delta K_{it}}$(K5)	−.279	.165	−.030	.43	−3.16	3.78	−.50
				$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\overline{R}^2$	$\bar{t}(\hat{\alpha}_1)$	$\bar{t}(\hat{\alpha}_2)$	$\bar{t}(\hat{\alpha}_3)$
K1	ΔK_{it}	$K_{i,t-1}, Q_{it}$	D_{it}(D1)	−.228	.090	.608	.30	−1.91	1.86	.41
K5	ΔK_{it}	$\Delta K_{i,t-1}, \Delta Q_{it}$	$\widehat{\Delta D_{it}}$(D1)	.339	.113	2.845	.17	1.60	1.59	.57

have the wrong sign (positive) and the cross-sectional distributions of *t*-statistics for these coefficients are in line with the hypothesis that these variables have little effect.

One might argue that the poor performance of the *2SLS* regressions just reflects the fact that the first stage regressions from which the instrumental variable estimates are obtained provide poor explanations of dividends and especially of investments. But *OLS* has been applied directly to equations (14)–(16), with again no improvement on the results for the corresponding models in Table 3. In fact, in predictions the *OLS* regressions on (14)–(16) do even worse in comparison with the models in Table 3 than the *2SLS* estimates. These results suggest that rather than being due entirely to inadequate instrumental variables, the poor performance of the *2SLS* regressions might in large part just reflect the general ineffectiveness of including any measure of investment as an explanatory variable in well-specified dividends models and any measure of dividends as an explanatory variable in well-specified investment models.

In short, there is no systematic evidence for the type of jointness or interdependence in the year-by-year dividend and investment decisions of firms that requires a simultaneous equations model. From the results presented above, single equation models of the dividend and investment decisions of firms seem appropriate.

This finding is in complete contradiction to the results of Dhrymes and Kurz, and we can suggest some reasons for the differences. First, although their theory is more consistent with time-series models applied to individual firms, and although their basic data are time-series for 181 firms, they choose to fit their dividend and investment equations with cross-sectional regressions in which the parameters are reestimated yearly. This forces the coefficients of explanatory variables to be the same for all firms. There is evidence in their own results that this is a misspecification. Using dummy variables, they allow fitted constants to vary by industry and observe significant differences. If dividends are correlated with other explanatory variables in their investment equation, then including dividends in the investment regressions may just be a way to adjust in part for differences among firms in the coefficients of other variables. A similar phenomenon may arise when investment is included in the cross-sectional dividend regressions.

Any problems due to differences in true coefficients among firms might be exacerbated by the fact that Dhrymes and Kurz use sales to deflate dividends and investments, and generally deflate variables which they treat as exogenous, like profits, with capital stock. The relationships among such ratios may differ more across firms and industries than would relationships among raw measures of the variables. And the use of such deflators raises the more basic issue of whether estimating relationships among the ratios sheds any light on the theoretical dividend and investment models that they propose to test.

Finally, empirical studies of dividend and investment models consistently find that last period's dividend is important in explaining this period's dividend, and that last period's capital stock is important in explaining this period's investment. Dhrymes and Kurz omit these lagged variables from their published regressions. But there can be correlation across firms between paired values of dividends and capital stock (or investment) that does not imply interdependence between the dividend and investment decisions of individual firms. For example, Dhrymes and Kurz probably induce such correlation by deflating dividends and investments with sales. Thus the dividend variable in their investment equation might in part be a proxy for the omitted lagged capital stock, while the investment variable in the dividend equation proxies for the omitted lagged dividend.

The omission of appropriate lagged variables, the inappropriateness of constraining the parameters of dividend and investment models to be the same for all firms, and the choice of deflators may account for another of the Dhrymes-Kurz findings that does not show up in this study. In their cross-sectional regressions, *2SLS* dividend residuals are highly correlated (usually .7 or more) with *2SLS* investment residuals. In our time-series regressions the cross-sectional distributions of the correlations between the time-series of *2SLS* dividend and investment residuals for individual firms are symmetric about zero and quite consistent with the hypothesis of no relationship. Indeed the distributions are like those for the apparently more relevant single equation models which are discussed next.

IV. Relationships Between the Parameters and Between the Residuals of Dividend and Investment Models

The last step in the search for interdependence between the dividend and investment decisions of firms is to examine the correlations between the *OLS* residuals of single equation dividend and investment models and between various parameter estimates. Table 7 shows the mean, standard deviation, and five fractiles of the cross-sectional distributions of correlation coefficients (computed firm-by-firm for each of the 298 firms in the sample) between the time-series of residuals from the Lintner dividend model **D1** and the time-series of residuals from the two-variable investment models **K1** and **K5**. Table 8 then shows the correlations of average dividend payout ratios (APO_i) and regression estimates of target payout ratios (TPO_i) and speed-of-adjustment coefficients (SOA_i) from the dividend model **D1** with average capital-to-output ratios (AKQ_i) and with regression estimates of investment speed-of-adjustment coefficients and target capital-to-output ratios (TKQ_i) from the investment models **K1** and **K5**.

There is no evidence in Table 8 that the

TABLE 7—CROSS-SECTIONAL DISTRIBUTIONS OF CORRELATIONS BETWEEN THE RESIDUALS OF THE DIVIDEND MODEL D1 AND THE INVESTMENT MODELS K1 AND K5

	D1 with K1	D1 with K5
Mean	−.034	.011
Standard Deviation	.252	.245
5 Fractiles		
.10	−.379	−.322
.30	−.145	−.119
.50	−.026	.022
.70	.119	.145
.90	.303	.331

TABLE 8—CORRELATIONS OF ESTIMATED DIVIDEND AND INVESTMENT PARAMETERS

	Dividend Parameter		
Investment Parameter	SOA_i	TPO_i	APO_i
SOA_i of Model **K1**	.037	.003	.004
SOA_i of Model **K5**	.082	.098	.029
TKQ_i of Model **K1**	−.056	.025	.023
TKQ_i of Model **K5**	−.103	.053	−.029
AKQ_i	−.099	.050	−.001

parameter estimates from the dividend regression model **D1** are correlated with parameter estimates from investment regression models. The correlations of the estimated speed-of-adjustment coefficients (SOA_i) and target payout ratios (TPO_i) of the dividend model with the speed-of-adjustment coefficients (SOA_i) and target capital-to-output ratios (TKQ_i) of the investment models are all small. And there is no evidence in Table 7 of systematic relationships between the residuals of the dividend model and the residuals of the investment models. The means of the correlations between dividend and investment residuals are close to zero, and the cross-sectional distributions are more or less symmetric about zero.

V. Conclusions

Modigliani and Miller show that in a perfect capital market, period-by-period investment decisions by a firm are separable from its dividend decisions. The results of this paper are consistent with this view. Whatever imperfections are present in the capital market are not sufficient to cause our data to reject the hypothesis that there is a rather complete degree of independence between the dividend and investment decisions of firms. Specifically there is no evidence for the type of jointness in dividend and investment decisions that would require treating them, via simultaneous equations models, as interdependent endogenous variables. And there is also no evidence of systematic relationships between parameter estimates of dividend and investment models or of any relationships between the residuals of these models.

It is, however, well to end by noting that the coefficient estimates and residuals from the regressions contain sampling errors that could obscure relationships either between the true parameters or between the disturbances of the dividend and investment models. Thus the appropriate conclusion is that the hypothesis of no period-by-period association between the dividend and investment decisions of firms is not rejected by the data of this study.

REFERENCES

J. A. Brittain, *Corporate Dividend Policy*, Washington 1966.

H. B. Chenery, "Overcapacity and the Acceleration Principle," *Econometrica*, Jan. 1952, *20*, 1–28.

P. Dhrymes and M. Kurz, "Investment, Dividends, and External Finance Behavior of Firms," in R. Ferber, ed., *Determinants of Investment Behavior*, New York 1967.

R. Eisner, "A Permanent Income Theory for Investment: Some Empirical Explanations," *Amer. Econ. Rev.*, June 1967, *57*, 363–90.

E. F. Fama and H. Babiak, "Dividend Policy: An Empirical Analysis," *J. Amer. Statist. Assn.*, Dec. 1968, *63*, 1132–61.

——— and M. H. Miller, *The Theory of Finance*, New York 1972

J. P. Gould and R. N. Waud, "The Neoclassical Model of Investment Behavior: Another View," *Int. Econ. Rev.*, forthcoming.

D. W. Jorgenson, "Econometric Studies of Investment Behavior: A Survey," *J. Econ. Lit.*, Dec. 1971, *9*, 1111–47.

———, "Rational Distributed Lag Functions," *Econometrica*, Jan. 1966, *34*, 135–49.

——— and C. Seibert, "A Comparison of Alternative Theories of Corporate Investment Behavior," *Amer. Econ. Rev.*, Sept. 1968, *58*, 681–712.

L. M. Koyck, *Distributed Lags and Investment Analysis,* Amsterdam 1954.

J. Lintner, "Distribution of Incomes of Corporations among Dividends, Retained Earnings and Taxes," *Amer. Econ. Rev.*, May 1956, *46*, 97–113.

T. Mayer, "Plant and Equipment Lead Times," *J. Bus., Univ. Chicago,* Apr. 1960, *33*, 127–32.

J. R. Meyer and E. Kuh, *The Investment Decision,* Cambridge 1957.

M. H. Miller and F. Modigliani, "Dividend Policy, Growth, and the Valuation of Shares," *J. Bus., Univ. Chicago,* Oct. 1961, *34*, 411–33.

F. Modigliani and M. H. Miller, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261–97.

J. Tinbergen, "Statistical Evidence on the Acceleration Principle," *Econometrica,* May 1958, *5*, 164–76.

Standard Statistics Corporation, Compustat Tapes, New York 1946–68.

On the Impact of Uncertainty on the Value and Investment of the Neoclassical Firm

By GUY V. G. STEVENS*

Despite recent advances incorporating uncertainty into the theory of the firm, major problems remain which mar the theory and prevent its practical application. This paper explores and offers solutions to two such problems: the form of the neoclassical firm's objective function under uncertainty, and the derivation of equations for investment and factor inputs from the maximization of a properly chosen objective function.

The expression adopted in Section I as the best available objective function for the neoclassical firm is derived from the well-known Lintner-Mossin-Sharpe (LMS) market valuation equation (see the articles by John Lintner (1965a), Jan Mossin, and William Sharpe). However, the major point of this section is that the LMS equation, as presently written, unnecessarily limits the scope of the analysis of optimal firm decisions, particularly investment decisions. Implicit in the currently accepted version of the equation is the assumption that today's decisions affect firm cash flow only in the present period. Truly dynamic problems such as those attacked in the investment analysis pioneered by Robert Eisner and Robert Strotz cannot, therefore, be handled. The situation can be remedied, but we will see that this is more difficult in some cases than others. It will depend on one's choice of assumptions about the randomness of *future* values of the market parameters: the riskless rate of interest and the market price of risk.

In the second part of this paper, a revised version of the LMS valuation equation is used to derive optimal investment, labor input, and production policies for models that have been extensively studied under certainty, but which have been treated inadequately or not at all under uncertainty. In particular, we will cover 1) the standard neoclassical case where the firm faces constant factor costs and constant returns to scale, and 2) the Eisner-Strotz model, where the cost of investment goods is no longer constant, but a function of the rate of investment.

I. Equations for the Value of the Firm Under Uncertainty

A. *The Value of the Firm Under Conditions of Certainty*

For purposes of comparison, it might be helpful to note that in this section we seek an analog for conditions of risk or uncertainty to the equation for the market value of the firm under certainty. Under conditions of perfect financial markets, the latter is equal to the sum (or integral) of discounted net cash flows exclusive of financial costs $[X(t)]$.[1] Merton Miller and Franco

---

* Chief, Quantitative Studies, Division of International Finance, Federal Reserve Board. My thanks go to Richard Bower, Robert Mackay, and my colleagues Roger Craine, Michael Laub, Richard Porter, Peter Tinsley, and Peter von zur Meulen for helpful comments and constructive criticism. An earlier draft of this paper was presented at the 1972 meetings of the Southern Economic Association. Naturally I am solely responsible for the opinions and remaining errors in this version.

[1] Under these conditions, the famous Modigliani-Miller theorem (1958) will hold: the firm's financial policy will be unrelated to the value of the firm. It is shown in Appendix A to my 1973 paper that this

Modigliani (1961) have shown the first expression to be equal to a number of alternatives, among which is the sum of discounted dividends $[DIV(t)]$ over time. Thus, under certainty, the value of the firm, at any time 0, is equal to:

$$(1) \quad \int_0^\infty X(t)e^{-rt}dt = \int_0^\infty DIV(t)e^{-rt}dt$$

where r is the rate of interest (here, for convenience only, assumed constant for all t).

B. *The Lintner-Mossin-Sharpe Model and Its Competitors*

The model developed below is based on a generalization of the LMS asset valuation equation. The assumptions and structure of the theory underlying the LMS equation have been detailed and criticized in a number of recent articles.[2] Rather than restate this work, I will present the broad outlines of the LMS theory, going into detail only to argue that, at the present time, its virtues outweigh its defects.

As is well known, the LMS equation for the value of any given firm is determined by the equilibrium conditions established in a market of atomistic portfolio investors. With respect to the behavior of these investors, the assumptions of the *original* LMS model were as follows:

1. The utility function of each investor is a function of the expected value and variance of return from his portfolio; each investor is risk averse, i.e., all other things equal, less variance is preferred to more.
2. A perfect market exists for riskless borrowing and lending.
3. All investors have a one-period horizon.
4. All investors have identical estimates of expected returns and variance for each opportunity.

Subsequently, the last three assumptions have been relaxed without changing the basic form of the implied equation for the value of the firm.[3]

The major weakness of the model is generally agreed to be the first assumption: the dependence of the investor's utility on expected return and variance alone. It is well known that the mean-variance approach is consistent with the widely accepted theory of expected utility maximization only if the investor's utility function is quadratic or all probability distributions are normal. Quadratic utility functions are rejected almost universally because of their unrealistic assumptions (see, for example, Lintner (1965a, fn. 20)); more people are willing to assume normality as an approximation to reality and Eugene Fama (1971) has gone somewhat further by showing that for the class of stable symmetric distributions very similar equations for the value of the firm result. However, I am sure that many will consider even Fama's relaxed version of assumption 1 to be too restrictive.

One can applaud efforts to work with more general assumptions and at the same time defend efforts to develop the implications of the LMS and related models. As far as the subject matter of this paper is concerned, the defense is that a theory of the firm based on the LMS valuation equation has a set of desirable properties that is not shared by any other existing model.

First, the LMS model has the essential property that the value of the firm is determined by the portfolio decisions of investors in a competitive securities market.

---

theorem holds for the models developed in the present paper under conditions of uncertainty. Thus we can ignore leverage.

[2] See Lintner (1965a), Robert Litzenberger and Alan P. Budd, and Michael Rothschild and Joseph Stiglitz.

[3] See Lintner (1965b) for the relaxation of assumption 2 and Eugene Fama (1970, 1971) for the relaxation of 3 and 4.

Hence managers who act to maximize the firm's value according to the LMS equation are following the neoclassical model: maximizing the *market* value of the firm. The study of the neoclassical model is of more than theoretical interest, since it provides, at present, our best description of corporate investment activity at all levels of data aggregation (see the articles by Charles Bishoff, and Dale Jorgenson and Calvin Siebert).[4]

As far as I know, with the exception of Jack Hirshleifer's contributions—which have their own drawbacks[5]—no other theory of the firm under uncertainty has forged the link between the firm's market value and its objective function. Interesting results have been derived by authors assuming a variety of alternative objectives (see Edward Zabel and Rothschild and Stiglitz for expected profit maximization); yet no one has shown how any of these alternatives is related to either the maximization of the firm's market value or some other theory of entrepreneurial behavior.

The second virtue of the LMS framework is that the firm is not artificially constrained in its investment policy with regard to the risk characteristics of the investments it selects. Its investments need not be limited to those in a given "risk class," however that may be defined. Nor are the firm's investments limited to those that do not affect the overall risk of its profit stream, as is the assumption of the Simon-Theil certainty equivalent approach (see Henri Theil, p. 349).

C. *The Lintner-Mossin-Sharpe Valuation Equation*[6]

Despite this impressive list of credits, the LMS equation in its present form does have a serious drawback. It may be general with respect to the risk characteristics that it can handle, but it is severely limited with respect to the *time* characteristics of the decisions it can analyze. The only decisions which can be incorporated into the LMS equation as it now stands are those whose effects are limited in time to the current period. The cause lies with the treatment of capital gains.

Let us start with a careful description of Lintner's version of the valuation equation (1965a, pp. 26–33). Let $Vi(0)$ be the aggregate market value of the equity of the ith company at time 0. Lintner shows in his world of risk-averse stockholders that:

$$(2)\quad Vi(0) = [1/r(0)]\left\{E[Ri(0)] - m(0)\left[\mathrm{Var}\,(Ri(0)) + \sum_{j=1}^{n} \mathrm{Cov}\,(Ri(0), Rj(0))\right]\right\}$$

where

$Ri(0)$ = the total return on the equity of company i at time 0, equal to cash dividends and capital gains. $E(Ri(0))$, Var $(Ri(0))$, and Cov $(Ri(0), Rj(0))$ are the moments of $Ri(0)$: its expected value, variance, and covariance with the return of any other firm, j.

---

[4] Traditionally the maximization of the value of the firm has also been supported as a normative criterion, leading under usual conditions to the maximization of the utility of the firm's stockholders. However, recently Stiglitz, and Michael Jensen and John Long have independently demonstrated that, under mean-variance uncertainty, this normative conclusion rarely follows. Nevertheless, most of the results derived below concerning the market value of multiperiod investments and optimal investment and labor policies should carry over, qualitatively unchanged, to the maximization of that objective function that does lead to maximum social welfare under uncertainty.

[5] Hirshleifer's model, using the time-state preference framework, possesses all the desirable properties claimed for the LMS theory; however, it too has drawbacks, the seriousness of which only time can determine: 1) a usual assumption is that all future states of the world can be defined by a finite set of states, a separate market existing for claims on each state; the latter, at least, is clearly not true; 2) most agree, including Hirshleifer, p. 277, that it is very difficult, if not impossible, to use the model empirically—because of the difficulty of relating observed security data, which relate to many states and times, to unobserved prices for claims in individual states.

[6] In this section, we will stick to the original form of the LMS equation. It should not be difficult to generalize the results below for the valuation equations derived from the more general models mentioned above.

$r(0)$ = the riskless rate of interest at time 0.

$m(0)$ = the market price of risk at time 0.

LMS and numerous other writers have studied how this equation can be used to derive the firm's optimal capital budget.[7] However, virtually no one has distinguished between a given investment's possible impact on the separate components of $R_i(0)$, dividends and capital gains—or, alternatively, current and future cash flows. Even though cases have been considered where an investment may affect cash flow over many periods, we have never been shown how to calculate the impact of these future effects on the value of the firm.

The one possible exception to the last statement is a recent article by Lintner (1971) where he attempts, as I do below, to handle the capital gains term by integrating it out. However, as discussed in the next section, his solution for the value of the firm is not a solution to the basic LMS valuation equation (2), above.[8]

The inability to handle multiperiod investments is caused by the fact that LMS have given us in equation (2) a *difference* or *differential* equation in $Vi(0)$, rather than a solution for $Vi(0)$ in terms (only) of its underlying determinants—dividends or cash flow over time. We shall derive these solutions in the next two sections.

D. *Case I: Future Riskless Interest Rates and Market Prices of Risk Known with Certainty*[9]

Let us first reformulate Lintner's equation (2) by breaking up $Ri(0)$ into dividends $[DIVi(0)]$ and capital gains $[\dot{V}i(0)$ or in period form, $Vi(1)-Vi(0)]$. The continuous form of equation (2) is given in (3). For the difference equation version of (3) one substitutes $Vi(1)-Vi(0)$ for $\dot{V}i(0)$.

The solutions to the differential or difference form of equation (3) depend on further assumptions to be made about the stochastic nature of $r(t)$ and $m(t)$. If we assume that the stockholder knows *future* riskless interest rates and market prices of risk with *certainty*, then a solution to (3) is readily available. For the continuous case, consider the possible solution shown as equation (4). For this solution, note that for each time s, including $s=0$, $\dot{V}i(s)$ is given by equation (5). Since $r(s)$ and $m(s)$ are known with certainty for all future time, there are no random variables in $\dot{V}i(s)$. Hence $E(\dot{V}i(0))$ equals simply $\dot{V}i(0)$ and Var $(\dot{V}i(0))$ equals zero; this holds for any other time (s) as well. Since $\dot{V}i(0)$ and $\dot{V}j(0)$ are nonstochastic, Var $(DIVi(0)+\dot{V}i(0))$ in equation (3) equals Var $(DIVi(0))$, and likewise, Cov $(DIVi(0)+\dot{V}i(0),\ DIVj(0)+\dot{V}j(0))$ equals Cov $(DIVi(0), DIVj(0))$.

After substituting equations (4) and (5) into the differential equation (3), and noting that for solution (4) $E(\dot{V}i(0))$ equals $\dot{V}i(0)$, equation (3) becomes: $m(0)[\text{Var}\ (DIVi(0))+\sum_j \text{Cov}\ (DIVi(0), DIVj(0))]$ equal to

$$m(0)[\text{Var}\ (DIVi(0)+\dot{V}i(0))+\sum_j \cdot\text{Cov}\ (DIVi(0)+\dot{V}i(0),\ DIVj(0)+\dot{V}j(0))]$$

I noted in the last sentence that these terms are equal at any time including 0, because of the nonrandomness of $\dot{V}i(0)$ and $\dot{V}j(0)$. Thus the differential equation (3) reduces to zero and equation (4) is a solution.

---

[7] See Robert Hamada and Fama (1971).

[8] See fn. 10 below.

[9] It becomes clear near the end of this section that a further necessary condition of Case I is that the moments at the time of decision, 0, of dividends for $t>0$ are not affected by random events after time 0.

$$Vi(0) = \frac{1}{r(0)}\left\{E[DIVi(0) + \dot{V}i(0)] - m(0)\cdot\text{Var}\,[DIVi(0) + \dot{V}i(0)] - m(0)\left[\sum_j \text{Cov}\,(DIVi(0) + \dot{V}i(0),\ DIVj(0) + \dot{V}j(0))\right]\right\} \tag{3}$$

$$Vi(0) = \int_0^\infty \left\{E[DIVi(\text{t})] - m(\text{t})\left[\text{Var}\,(DIVi(\text{t})) + \sum_j \text{Cov}\,(DIVi(\text{t}),\ DIVj(\text{t}))\right]\right\} \cdot \exp\left\{-\int_0^{\text{t}} r(x)dx\right\} dt \tag{4}$$

$$\dot{V}i(s) = r(s)Vi(s) - E(DIVi(s)) + m(s)\left[\text{Var}\,(DIVi(s)) + \sum_j \text{Cov}\,(DIVi(s),\ DIVj(s))\right] \tag{5}$$

$$Vi(0) = \sum_{\text{t}=0}^{\infty}\left\{E[DIVi(\text{t})] - m(\text{t})\left[\text{Var}\,(DIVi(\text{t})) + \sum_j \text{Cov}\,(DIVi(\text{t}),\ DIVj(\text{t}))\right]\right\} \cdot \prod_{p=0}^{\text{t}} [1 + r(p)]^{-1} \tag{6}$$

The solution to the difference equation analog to (3) is quite similar, and is given in (6).

In summary, for Case I the value of the firm under uncertainty is the infinite sum of risk-adjusted expected dividends, discounted by the riskless rate of interest. This expression, it should be emphasized, is very different from the well-known equation for the value of the firm which is the fruit of the constant risk class theory criticized above (see Section IB above). This latter states that the value of the firm is equal to the infinite sum of expected dividends (unadjusted for risk) discounted by a risk-adjusted rate of interest (the cost of capital for all firms in the given risk class).[10]

---

[10] This is an appropriate place to discuss Lintner's proposed solution to his similar, but not identical, model mentioned in the previous section. I do not want to take a position on whether the proposed Lintner equation is a valid solution to his new model: that would require a lengthy discussion of the equilibrium equations of that model and his method of proof, a discussion that is quite beyond the scope of this article. However, it is appropriate to show that his solution is *not* a solution to the model considered in this article, the one defined by the valuation equation (2). Attention will be limited to the only case Lintner considers, the special case where future riskless rates of interest or discount rates are known with certainty.

Lintner's proposed solution is formally identical to those proposed by proponents of the constant risk class hypothesis; i.e., uncertainty is fully accounted for by adjusting the discount rate for risk. Lintner (1971, p. 194) proposes the following solution:

$$V(0) = \int_0^\infty D(0)\exp\,[\bar{g}_D - \alpha_M C\sigma^2(g_D)]\text{t}\exp\,[-r\text{t}]dt$$

$V(0)$, $D(0)$, are, as before, the value of the firm and dividends at time 0; r is the riskless rate of interest (or discount rate) assumed for convenience only to be constant over time; $\bar{g}_D$ is the expected growth rate of dividends over time and $\sigma^2(g_D)$ is the variance of this growth rate; α_M and C are market parameters; covariances with other returns or growth rates are ignored, again for convenience only. Lintner's solution, like the above one for the special Case I, implies that the variance of capital gains is zero; taking this into account, The most interesting point about solutions (4) and (6)—and probably quite unrealistic—is that capital gains are nonstochastic. Although they are likely to be nonzero, in this world the capital gains of the firm are known with certainty. This result depends on the special assumptions of Case I. If we let future one-period riskless rates become random variables and/or take a Baysian approach and assume that the expected value and variance of dividends are random, then capital gains become random. This is the major problem of the next section.

E. *Case II: The Riskless Rate and Market Price of Risk for Future Periods as Random Variables: The General Case*

The first and most important thing to note about the general case is that capital gains no longer are known with certainty. Capital gains are $V(1)-V(0)$, and the difference equation version of (3) shows that $V(1)$ is a function of $r(1)$ and $m(1)$, which are random variables at time 0, the time of decision. Hence the solution to Case I no longer holds.

Although the general solution to the basic difference or differential equation (3) is much more complex, its structure can be made clear: the solution to the general case is constructed from the solution to the special case (equation (6), above), by 1) substituting the *expected values* of future riskless interest rates, etc. for the previously known values; and 2) subtracting another series of risk discounts, reflecting the variance of capital gains. Here we shall limit our attention to the solution of the general difference equation.[11] Ignoring the covariance terms which are irrelevant to the question at hand, the difference equation version of (3) becomes:

$$V(0) = \frac{1}{r(0)} \{E[DIV(0)] + E[V(1) - V(0)] - m(0) \operatorname{Var} [DIV(0) + V(1) - V(0)]\}$$

Since at time 0 (and any time t in the future), the value of the firm at that time is known with certainty, $E(V(0)) = V(0)$ and $\operatorname{Var} (V(0)) = 0$. We will assume that the value of the firm in any period will not be a function of the returns (DIV) in earlier periods;[12] thus $DIV(0)$ and $V(1)$ will be

---

along with the inconsequential neglect of covariances, implies that any proposed solution must satisfy the modified LMS differential equation: $r(s)V(s) = E[DIV(s)] + \dot{V}(s) - m(s) \operatorname{Var} [DIV(s)]$. Unless I am mistaken, Lintner's proposed solution does not satisfy the above equation. First, it is impossible to find anything in the solution that corresponds exactly to the moments $E[DIV(s)]$ and/or $\operatorname{Var} [DIV(s)]$; this is probably due to the fact that Lintner is concentrating on equilibrium growth rates and, hence, lognormal distributions. Thus: $E[DIV(s)] = E[D(0)e^{gs}]$ or $E[D(0) \cdot \exp \{\int_0^s g(v)dv\}] \neq D(0)e^{\bar{g}s} = \exp \{E[lnD(s)]\}$. On the other hand, assuming we can approximate $E[DIV(s)]$ by $D(0)e^{gs}$, we can then rewrite Lintner's solution as equal to the integral of expected dividends discounted by a risk-adjusted discount rate:

$$V(s) = \int_s^\infty E[DIV(\mathrm{t})] \exp \{-\alpha_M C\sigma^2(g_D)(\mathrm{t} - s)\} \cdot \exp \{-r(\mathrm{t} - s)\} d\mathrm{t}$$
$$= \int_s^\infty E[DIV](\mathrm{t})] \exp \{-[r + \alpha C\sigma^2(g_D)](\mathrm{t} - s)\} d\mathrm{t}$$

Substituting the market price of risk (m) for Lintner's similar market parameters $\alpha_M C$ and differentiating Lintner's solution with respect to s yields: $rV(s) = E(DIV(s)) + \dot{V}(s) - m\sigma^2(g_D)V(s)$. Lintner's proposed solution satisfied the LMS differential equation only where $\operatorname{Var} (DIV(s))$ equals $\sigma^2(g_D)V(s)$; i.e., the variance of the dividend stream is proportional to the value of the firm. This is not generally true, not even for the special class of models where dividends (or expected dividends) show a steady growth rate.

[11] The solution to the general differential equation requires added assumptions concerning the moments of the random variables to ensure, among other things, the continuity of $Vi(0)$. Probably as $t_0 \to t_1$, the permissible variance in the riskless rate of interest and all other random variables during the small period, $t_1 - t_0$, must approach zero.

[12] Since $V(1)$ is a function of the expected value and variance of dividends in periods 1 to ∞, this means that we are assuming that the outcome for dividends in period zero doesn't affect the moments of dividends for future periods. Such interdependence or serial correlation can be built into the model by adding the covariance between $DIV(0)$ and $V(1)$, but this would further complicate an already very complicated expression, equation (10) below.

$$(8)\quad V(0) = \frac{1}{1+r(0)}\left\{E[DIV(0)] - m(0)\,\mathrm{Var}\,[DIV(0)] + E\left(\frac{1}{1+r(1)}\right)E[DIV(1)] - E\left[\frac{m(1)}{1+r(1)}\right]\mathrm{Var}\,[DIV(1)]\right\} - \frac{m(0)}{1+r(0)}\left\{\mathrm{Var}\left[\frac{1}{1+r(1)}\right](E[DIV(1)])^2 + \mathrm{Var}\left[\frac{m(1)}{1+r(1)}\right](\mathrm{Var}\,[DIV(1)])^2 - 2\,\mathrm{Cov}\left[\frac{1}{1+r(1)}, \frac{m(1)}{1+r(1)}\right] \cdot E[DIV(1)]\,\mathrm{Var}\,[DIV(1)]\right\}$$

statistically independent and $\mathrm{Var}\,[DIV(0) + V(1) - V(0)] = \mathrm{Var}\,[DIV(0)] + \mathrm{Var}\,[V(1)]$. Hence the basic difference equation simplifies to:

$$(7)\quad V(0) = \left(\frac{1}{1+r(0)}\right)\{E[DIV(0)] + E[V(1)] - m(0)\,\mathrm{Var}\,[DIV(0)] - m(0)\,\mathrm{Var}\,[V(1)]\}$$

For expository purposes, it is best to start the analysis with a firm (asset) which has a *finite* life. In such a case one can derive the value of the firm at time 0 by backward recursion from the last period n. The valuation expression for the asset with finite life, n, leads naturally to the general equation for the value of an asset with an infinite life.

Consider the valuation of the simplest multiperiod investment, an investment yielding returns over 2 periods: $DIV(0)$, $DIV(1)$.[13] For this example, since there are no returns after period 1, the value of the asset is zero in all periods greater than 1. Thus there is no uncertainty concerning capital gains during period 1. Using equation (7) we know therefore that:

$$V(1) = \frac{1}{1+r(1)}\{E[DIV(1)] - m(1)\,\mathrm{Var}\,[DIV(1)]\}$$

To determine $V(0)$, according to equation (7) we need to know $E[V(1)]$ and $\mathrm{Var}\,(V(1)]$. From the preceding:

$$E[V(1)] = E\left(\frac{1}{1+r(1)}\right)E[DIV(1)] - E\left(\frac{m(1)}{1+r(1)}\right)\mathrm{Var}\,[DIV(1)]$$

$$\mathrm{Var}\,[V(1)] = \mathrm{Var}\left(\frac{1}{1+r(1)}\right)(E[DIV(1)])^2 + \mathrm{Var}\left(\frac{m(1)}{1+r(1)}\right)\cdot(\mathrm{Var}\,[DIV(1)])^2 - 2\,\mathrm{Cov}\left(\frac{1}{1+r(1)}, \frac{m(1)}{1+r(1)}\right)\cdot E[DIV(1)]\,\mathrm{Var}\,[DIV(1)]$$

Hence, for the value of the firm (asset) at time 0 we have equation (8). The last term in (8) is a risk discount to take account of the loss in value due to the riskiness of capital gains; the first term is nearly the same as the expression for the value of the firm under the assumptions of Case I (equation (6)), except that $1/[1+r(1)]$ and $m(1)/[1+r(1)]$ are replaced by their expectations.[14] It might be noted that if only future riskless rates of interest are random (and not also future market prices of risk), then the risk discount for capital gains simplifies considerably, becoming:

---

[13] We call these returns "dividends" to maintain consistency with the preceding sections; these returns can be thought of as stream of cash flow, as will be seen in Section II.

[14] It is worth noting that equation (8) provides us with the proper way under mean-variance uncertainty to derive the discounted present value of future returns. Thus the present value at time zero of dividends in period 1 is given by the last three lines of (8). For more details, see my 1973 paper, p. 16.

$$-\frac{m(0)}{1+r(0)}\operatorname{Var}\left[\frac{1}{1+r(1)}\right][E[DIV(1)] - m(1)\operatorname{Var}[DIV(1)]]^2$$

In similar fashion, one can use equation (8) to derive the expected value and variance of capital gains for the *three-period* firm, and, in conjunction with equation (7), proceed recursively to the value of the *n*-period firm.[15] In so doing, the principles emerge which lead to the solution of the general, infinite-period case. First, the *present value* of all risk discounts are carried along in the value expression; i.e., in the *n*-period case we have $2n-1$ terms reflecting discounts for risk: *n* terms measuring the present value (cost) due to the uncertainty of returns (dividends) in periods zero through $n-1$; and $n-1$ terms reflecting the present value caused by the uncertainty of capital gains in periods zero through $n-2$.

For each type of risk discount, its value *in the period in which the uncertainty occurs* is equal to the variance of the particular return in question (capital gain or dividend) times the ratio of that period's market price of risk and riskless rate of interest. The present value at time zero of each risk discount, which affects the value of the firm in time zero, is calculated in a fairly familiar way: we discount the variance in the future period *i* by the product of 1) the expected values of the discount factors $E[1/[1+r(j)]]$, $j=0,\ldots,i-1$, and 2), for the last period, $E[m(i)/[1+r(i)]]$.

Finally, it should be noted how the variance of capital gains changes as we move from two to three to *n* periods. In each period the variance of capital gains is made up of three terms, one multiplied by the variance of $[1/[1+r(i)]]$, one by the variance of $[m(i)/[1+r(i)]]$, and one by the negative of the covariance of these two variables. Both of the first two terms are a series of *embedded squares* (in the general case, an infinite series). The squared terms arise because the nonstochastic coefficient of a random variable is squared to form its variance; the squares are embedded because in calculating the variance of capital gains in time t, the random variables $1/(1+r(t+1))$ and $m(t+1)/(1+r(t+1))$ are multiplied by terms that are already squares—these reflecting the variance of capital gains in succeeding periods t+2, t+3, etc.

Now let us move on to the case of a firm or asset with an infinite life. To facilitate the writing of the general solution define the two infinite sums, $A(s)$ and $B(s)$ shown in (9). The terms in $A(\ )$ and $B(\ )$ on the right-hand side of equations (9) are of a higher index or order than the index *s* on the left-hand side. Hence by substituting for $A(s+1)$ and $B(s+1)$ in terms of the definition, we can express as many of the higher order *A* and *B* as we desire as functions of the moments of the dividend stream, the market parameters, and *A* and *B* of still higher order. Further, the general expression (10) below tells us that each $A(i)$ and $B(i)$ is finite in value, since it is assumed that the market value of the variance of capital gains is finite; therefore, since each term in $A(s+1)$ and $B(s+1)$ in the above definitions is *discounted* to the present (*s*), the magnitude of higher order terms will approach zero as their index increases.

Using $A(i)$ and $B(i)$ as defined above, the general expression for the market value at any time *s* of a firm of infinite life is given by (10). To demonstrate that (10) is indeed a solution to the basic difference equation (7), form the expression for $V(s+1)$ and take its expected value.

---

[15] The three-period solution is provided in my 1973 paper, p. 17.

$$
\begin{aligned}
(9)\quad A(s) = {} & E[DIV(s)] + E\left[\frac{1}{1+r(s+1)}\right]E[DIV(s+1)] - E\left[\frac{m(s+1)}{1+r(s+1)}\right] \\
& \cdot \operatorname{Var}[DIV(s+1)] + \sum_{i=s+2}^{\infty}\left\{\prod_{j=s+1}^{i-1} E\left[\frac{1}{1+r(j)}\right]\left(E\left[\frac{1}{1+r(i)}\right]E[DIV(i)]\right.\right. \\
& \left.\left. - E\left[\frac{m(i)}{1+r(i)}\right]\operatorname{Var}[DIV(i)]\right)\right\} - E\left[\frac{m(s+1)}{1+r(s+1)}\right] \\
& \cdot\left\{\operatorname{Var}\left[\frac{1}{1+r(s+2)}\right][A(s+2)]^2 + \operatorname{Var}\left[\frac{m(s+2)}{1+r(s+2)}\right][B(s+2)]^2\right. \\
& \left. - 2\operatorname{Cov}\left[\frac{1}{1+r(s+2)}, \frac{m(s+2)}{1+r(s+2)}\right]A(s+2)B(s+2)\right\} \\
& - \sum_{i=s+3}^{\infty} E\left[\frac{m(i-1)}{1+r(i-1)}\right]\prod_{j=s+1}^{i-2} E\left[\frac{1}{1+r(j)}\right]\left\{\operatorname{Var}\left[\frac{1}{1+r(i)}\right][A(i)]^2\right. \\
& \left. + \operatorname{Var}\left[\frac{m(i)}{1+r(i)}\right][B(i)]^2 - 2\operatorname{Cov}\left[\frac{1}{1+r(i)}, \frac{m(i)}{1+r(i)}\right]A(i)B(i)\right\} \\
B(s) = {} & \operatorname{Var}[DIV(s)] + \operatorname{Var}\left[\frac{1}{1+r(s+1)}\right][A(s+1)]^2 + \operatorname{Var}\left[\frac{m(s+1)}{1+r(s+1)}\right] \\
& \cdot [B(s+1)]^2 - 2\operatorname{Cov}\left[\frac{1}{1+r(s+1)}, \frac{m(s+1)}{1+r(s+1)}\right]A(s+1)B(s+1)
\end{aligned}
$$

$E[V(s+1)]$ contains every term of $[1+r(s)]V(s)$ except $E[DIV(s)]-m(s)$ Var $[DIV(s)]$ and the term beginning with $-m(s)$.

But we can show that this latter expression is exactly the variance of $V(s+1)$ at time s. First note that at time s, $V(s+1)$ is an additive function of two random variables: $1/[1+r(s+1)]$ and $-m(s+1)/[1+r(s+1)]$. So the general form of Var $[V(s+1)]$ will be

$$
\begin{aligned}
& \operatorname{Var}[1/(1+r(s+1))]C^2 \\
& \quad + \operatorname{Var}[m(s+1)/(1+r(s+1))]D^2 \\
& \quad - 2\operatorname{Cov}[1/(1+r(s+1)), \\
& \qquad\qquad m(s+1)/(1+r(s+1))]CD
\end{aligned}
$$

where C and D are two constants. In $V(s+1)$ the terms multiplied by Var $[1/[1+r(s+1)]]$ are precisely equal to $A(s+1)$, so C equals $A(s+1)$; similarly D equals $B(s+1)$. Hence the term multiplying $m(s)$ in $V(s)[1+r(s)]$, mentioned above, equals Var $[V(s+1)]$. For equation (10), therefore, we have the following:

$$
\begin{aligned}
& [1+r(s)]V(s) - E[V(s+1)] \\
& \qquad = E[DIV(s)] - m(s)\operatorname{Var}[DIV(s)] \\
& \qquad\quad - m(s)\operatorname{Var}[V(s+1)]
\end{aligned}
$$

Since this holds for any period s, including 0, equation (10) is a solution to the LMS difference equation (7).

F. *Optimal Decision Making for the Special and General Cases*

Under some circumstances, we may assume that the decisions of one period will not affect the decisions of later periods: for example, under the assumptions that cur-

$$
\begin{aligned}
(10)\quad [1+r(s)]V(s) ={}& E[DIV(s)] - m(s)\operatorname{Var}[DIV(s)] + E\left[\frac{1}{1+r(s+1)}\right] E[DIV(s+1)] \\
& - E\left[\frac{m(s+1)}{1+r(s+1)}\right]\operatorname{Var}[DIV(s+1)] \\
& + \sum_{i=s+2}^{\infty} \prod_{j=s+1}^{i-1} E\left[\frac{1}{1+r(j)}\right] \cdot \left\{ E\left[\frac{1}{1+r(i)}\right] E[DIV(i) \right. \\
& \left. - E\left[\frac{m(i)}{1+r(i)}\right]\operatorname{Var}[DIV(i)] \right\} \\
& - m(s)\left\{ \operatorname{Var}\left[\frac{1}{1+r(s+1)}\right][A(s+1)]^2 + \operatorname{Var}\left[\frac{m(s+1)}{1+r(s+1)}\right]\cdot[B(s+1)]^2 \right. \\
& \left. - 2\operatorname{Cov}\left[\frac{1}{1+r(s+1)}, \frac{m(s+1)}{1+r(s+1)}\right] A(s+1)E(s+1) \right\} \\
& - E\left[\frac{m(s+1)}{1+r(s+1)}\right]\left\{ \operatorname{Var}\left[\frac{1}{1+r(s+2)}\right][A(s+2)]^2 + \operatorname{Var}\left[\frac{m(s+2)}{1+r(s+2)}\right][B(s+2)]^2 \right. \\
& \left. - 2\operatorname{Cov}\left[\frac{1}{1+r(s+2)}, \frac{m(2)}{1+r(s+2)}\right] A(s+2)B(s+2) \right\} \\
& - \sum_{i=s+3}^{\infty} E\left[\frac{m(i-1)}{1+r(i-1)}\right] \prod_{j=s+1}^{i-2} E\left[\frac{1}{1+r(j)}\right] \cdot \left\{ \operatorname{Var}\left[\frac{1}{1+r(i)}\right][A(i)]^2 \right. \\
& \left. + \operatorname{Var}\left[\frac{m(i)}{1+r(i)}\right][B(i)]^2 - 2\operatorname{Cov}\left[\frac{1}{1+r(i)}, \frac{m(i)}{1-r(i)}\right] A(i)B(i) \right\}
\end{aligned}
$$

rent decisions on investment, etc. are fully reversible next period and that current decisions do not influence things such as the price of investment goods or labor. Under such conditions, which may involve both Cases I and II above, the firm's decisions at time 0 will not affect capital gains or dividends in any future period; hence decisions can be made one period at a time. Equations (4), (6), and (10) show that, for both Cases I and II, if decisions are fully reversible, the full effect of decisions at time 0 on the value of the firm will be measured by their effect on the term: $(E[DIV(0)]-m(0)\operatorname{Var}[DIV(0)])/(1+r(0))$. This is the case where the one-period investment criterion of Lintner and others is applicable. I would think that the set of assumptions justifying the one-period criterion would be particularly unrealistic for most real corporations, but on the other hand, it does underlie such popular and successful investment models as Jorgenson's.

In those cases where the above rule is not optimal, the optimal decision procedure will differ greatly between Cases I and II. For Case I, actual outcomes for $r(t)$, $m(t)$, and $DIV(t)$ will not affect future moments for these variables. In this case the optimal decisions for every period may be determined at time 0. However, if we take a Bayesian approach and assume that the moments of the probability distribution of $DIV(t)$ are affected by previous outcomes, then the firm's decision procedure is complicated by the necessity of adopting a sequential approach, i.e., dynamic programming, whereby decisions at any time are made a function of all previous outcomes.

In the *general* case, in the absence of full reversibility (even if we eschew the Bayesian approach), the firm's optimal decisions at any time t will depend on developments in interest rates, etc., from time 0 to t. Here the dynamic programming approach will be required under both Bayesian and non-Bayesian assumptions.

II. Some Implications for the Neoclassical Firm's Optimal Investment Policy Under Uncertainty

The rule for the manager of a neoclassical firm is to make decisions so as to maximize the firm's market value. In this section the emphasis will be on optimal investment decisions; particular attention will be focused on the generalization to condition of mean-variance uncertainty of models studied by Jorgenson, and Eisner and Strotz, models which have played an important role in recent advances in the theoretical and empirical analysis of investment.

Throughout this section, where decisions at different points in time are interdependent, we will adopt the assumptions of Case I above: that future riskless one-period rates of interest and market prices of risk are certain. Thus the firm's managers should maximize equation (4) for continuous time and equation (6) for discrete time.

A. *Substituting for Dividends*

Equations (4) and (6), it will be recalled, indicate the the value of the firm is a function of market parameters and the mean, variance, and covariances of all future dividends. Dividend policy, however, does not affect the value of the firm directly, but only indirectly as it mirrors the firm's production and investment policies. In order to study at first hand the decisions that directly affect the value of the firm, it is necessary, therefore, to substitute for dividends in terms of investment and production variables.

This is done by using the definition of cash flow. By making familiar simplifying assumptions, such as no change in current assets and no taxes, one arrives at the following substitution:[16]

$$(11)\quad \begin{aligned} DIV &= pQ - wL - qI \\ &= pQ - wL - q(\dot{K} + dK), \end{aligned}$$

p = the selling price of the product
w = the wage rate
q = the price of investment goods
Q = the quantity of final output produced
L = the quantity of labor input
K = the real capital stock
d = the constant rate of depreciation
I = real gross investment, which equals, $\dot{K} + dK$

Except for the neglect of taxes, the above is identical to the substitution used by Jorgenson and Siebert.

To study optimal investment and production policy we substitute the right side of (11) for DIV(t) in equations (4) and (6), and maximize either of these with respect to the variables over which the firm has command. Making this substitution in the continuous model (4), we arrive at the objective function for the neoclassical firm, as shown in (12).

B. *Price Uncertainty in the Standard Neoclassical Model*

Let us start the exploration of optimal firm decisions under uncertainty with a simple extension of Jorgenson's neoclassical model. Assume a perfectly competitive firm with all factor prices independent of firm decision variables, especially the rate

---

[16] See my 1973 paper for a fuller development of the alternatives.

$$
\begin{aligned}
(12)\quad Vi(0) = \int_0^\infty \Big\{ & E[p(\mathrm{t})Q(K(\mathrm{t}), L(\mathrm{t})) - w(\mathrm{t})L(\mathrm{t}) - q(\mathrm{t})(\dot{K}(\mathrm{t}) + dK(\mathrm{t}))] \\
& - m(\mathrm{t}) \operatorname{Var} [p(\mathrm{t})Q(\ ) - w(\mathrm{t})L(\mathrm{t}) - q(\mathrm{t})(\dot{K}(\mathrm{t}) + dK(\mathrm{t}))] \\
& - m(\mathrm{t}) \sum_j \operatorname{Cov} [p(\mathrm{t})Q(\ ) - w(\mathrm{t})L(\mathrm{t}) - q(\mathrm{t})(\dot{K}(\mathrm{t}) + dK(\mathrm{t})), \\
& p_j(\mathrm{t})Q_j(\ ) - w(\mathrm{t})L_j(\mathrm{t}) - q(\mathrm{t})(\dot{K}_j(\mathrm{t}) + dK_j(\mathrm{t}))] \Big\} \exp \left\{ - \int_0^{\mathrm{t}} r(x)dx \right\} dt
\end{aligned}
$$

of change of capital. For simplicity (although this is not essential), assume that this firm faces uncertainty only in its selling price, p_i; which has an expected value, $E(p_i)$, variance Var (p_i), and covariance with any other price j, Cov (p_i, p_j).

Thus, for a given point in time, the expected value of cash flow (the first line of equation (12)) becomes: $E(p_i)Q_i(\ ) - wL_i - q(\dot{K}_i + dK_i)$. The variance of cash flow is simply Var $(p_i)Q_i^2$; assuming that other firms also experience uncertainty only in their prices, the covariance between the cash flow of firms i and j is Cov $(p_i, p_j) \cdot Q_i(\ )Q_j(\ )$.

If these substitutions are made into equation (12), only one term remains in $\dot{K}(\mathrm{t})$; as is well known,[17] the remaining $\dot{K}(\mathrm{t})$ can be integrated out, with $-q(\dot{K}+dK)$ being replaced by $-q(-\dot{q}/q+r+d)K$. With all terms in $\dot{K}(\mathrm{t})$ eliminated, we see that decisions and returns at any time t will be independent of returns and decisions at all other times; we can avoid the use of the calculus of variations and may minimize (12) simply by choosing K_i and L_i at each point in time so as to maximize that time's expected cash flow adjusted for risk:

$$
\begin{aligned}
(13)\quad & E(p_i)Q_i(\ ) - wL_i - q(-\dot{q}/q + r + d)K_i \\
& - m\Big[\operatorname{Var} (p_i)Q_i^2(\ ) \\
& + \sum_{j \neq i} \operatorname{Cov} (p_i, p_j)Q_i(\ )Q_j(\ ) \Big],
\end{aligned}
$$

Maximizing (13) with respect to K_i and L_i at t, the first-order conditions are, respectively,

$$
\begin{aligned}
(14)\quad & E(p_i)\partial Q_i/\partial K_i - q(-\dot{q}/q + r + d) \\
& - 2m \operatorname{Var} (p_i)Q_i(\partial Q_i/\partial K_i) \\
& - m \sum_{j \neq i} \operatorname{Cov} (p_i, p_j)Q_j\partial Q_i/\partial K_i = 0
\end{aligned}
$$

$$
\begin{aligned}
(15)\quad & E(p_i)\partial Q_i/\partial L_i - w \\
& - 2m \operatorname{Var} (p_i)Q_i\partial Q_i/\partial L_i \\
& - m \sum_{j \neq i} \operatorname{Cov} (p_i, p_j)Q_j\partial Q_i/\partial L_i = 0
\end{aligned}
$$

The first implication to notice is that, despite the risk in the system, which is taken into account in the maximization equations, factor proportions are the same for any given level of production as under perfect certainty. This can be seen by observing that (14) and (15) imply that $(\partial Q/\partial K)(\partial Q/\partial L) = q(-\dot{q}/q+r+d)/w$, the same condition as for cost minimization under certainty. This is not surprising, since there is no uncertainty associated with production as opposed to that in selling the product once produced.

Another, perhaps more important, implication is that constant returns to scale in production is no longer a bar to a determinate profit maximization equilibrium for the firm. Even with constant returns to scale—and probably with moderate increasing returns to scale as well—there is an optimum finite scale for the firm at every point in time. The reason is that the risk associated with the firm's profits increases nonlinearly with the *square* of the

---

[17] See my 1973 paper, Appendix B.

$$(16)\qquad K^* = \frac{aA\left[E(p_i) - m \sum_{j\neq i} \mathrm{Cov}\,(p_i, p_j)Q_j\right]\left\{\dfrac{q(-\dot{q}/q+r+d)(1-a)}{wa}\right\}^{(1-a)} - q(-\dot{q}/q+r+d)}{2m\,\mathrm{Var}\,(p_i)aA^2\left\{\dfrac{q(-\dot{q}/q+r+d)(1-a)}{wa}\right\}^{2(1-a)}}$$

$$(17)\qquad L^* = [(1-a)/a][q(-\dot{q}/q+r+d)/w]K^*$$

quantity produced; since the firm's risk has a cost attached to it (m), under uncertainty the cost function of the perfectly competitive firm is non-linear.[18]

In the Appendix, I consider the linear-homogeneous Cobb-Douglas case in some detail, deriving the optimal levels of capital and labor at t and showing that the sufficient conditions for a maximum are satisfied. As derived there the optimal levels of capital (K^*) and labor (L^*) for the Cobb-Douglas case turn out to be given by (16) and (17) (where a and A are parameters of the production function). As expected, (16) and (17) show that an increase in either type of risk, Var (p_i) or Cov (p_j, p_i), leads to *decreases* in both K^* and L^*.

C. *A Model with Temporally Dependent Decisions: The Eisner-Strotz Model under Mean-Variance Uncertainty*

Decisions at different points in time are interrelated in the Eisner-Strotz model because the cost of investment goods is assumed to be a function of the rate of investment. Here we will take the original Eisner-Strotz model and generalize its results for the case of mean-variance uncertainty. Again, as in the preceding example, the simplest kind of price uncertainty will be assumed.

As discussed above we shall assume that future riskless rates of interest and market prices of risk are known with certainty. Substituting for dividends, the value of the firm is determined by equation (12) above. However, the added assumptions of the Eisner-Strotz model lead to a special form of equation (12). The substantive assumption is, of course, that the cost of a unit of investment goods is a function of the rate of investment (I). In the particular version we shall follow (corresponding to their case I, pp. 79–84), a quadratic cost function for investment goods is adopted: $C(I) = kI + \gamma I^2$, where $C(I)$ is the total cost of a given level of investment.

Some differences between the Eisner-Strotz model and that underlying the original equation (12) are the following:

1) Labor costs are ignored; assuming that labor is optimally adjusted and the production function remains linear-homogenous, output becomes a function of capital alone, say, $Q = AK$.

2) The riskless rate of interest is assumed constant throughout the future: thus the discount factor in (12) becomes e^{-rt}. As well, the depreciation rate d is assumed equal to zero.

In order to have a determinate long-run equilibrium for their model, Eisner and Strotz assumed that marginal revenue exclusive of investment costs is a decreasing function of plant size (K), i.e., total revenues exclusive of investment costs are equal to $\alpha K - \beta K^2$. This condition is not necessary for purposes of long-run equilibrium under mean-variance uncertainty, but it will be adopted here to keep the present model as close as possible to the original. Eisner and Strotz do not explain how this decreasing marginal revenue comes about; for convenience we shall assume it is due to a negatively sloped demand curve. Let the firm's product price (p) be determined by the following demand curve: $p_i = a_i - b_iQ_i + u_i$; a_i and b_i are constants, and u_i is a random variable with mean zero, variance Var (u_i), and covariance with any other price Cov (u_i, u_j). Where convenient, the subscripts i will be dropped. This is the simplest kind of uncertainty of price—uncertainty independent of the level of price—but other kinds can be handled with only slightly more difficulty.

---

[18] The referee pointed out that Agnar Sandmo derives this result for the case where the firm's objective is to maximize expected utility.

Above it was assumed that output is proportional to the capital stock: $Q_i = A_iK_i$. Therefore total revenue exclusive of investment or capital costs is:

$$\begin{aligned} pQ &= (a_i - b_iA_iK_i + u_i)A_iK_i \\ &= aAK - bA^2K^2 + AKu \end{aligned}$$

This corresponds to $pQ(K, L) - wL$ in equation (12). In order that our symbols may correspond to those of Eisner and Strotz, we will let aA equal α and bA^2 equal β.

Putting the pieces together, in this model net cash flow *inclusive* of investment costs is: $\alpha K - \beta K^2 + AKu - k\dot{K} - \gamma(\dot{K})^2$. Except for the random variable AKu, this is the same as the Eisner-Strotz expression. The expected value of the firm's net cash flow equals: $\alpha K - \beta K^2 - k\dot{K} - \gamma(\dot{K})^2$. The variance of this cash flow is simply A^2K^2 Var (u) and the covariance between the cash flow of firms i and j is $A_iA_jK_iK_j$ Cov (u_i, u_j).

Making all the above substitutions into equation (12), and assuming for convenience that m(t) is a constant m, the value of the Eisner-Strotz firm becomes:

$$(18)\quad V_i(0) = \int_0^\infty \Big\{ [\alpha K - \beta K^2 - k\dot{K} - \gamma(\dot{K})^2] - m\Big[A^2K^2 \text{ Var } (u_i) + AK \sum_j A_jK_j \text{ Cov } (u_i, u_j) \Big] \Big\} \cdot \exp\{-rt\}\, dt$$

Except for the market parameters m and r, all unsubscripted variables refer to firm i. This equation differs from the original Eisner-Strotz formulation only in the addition of the variance and covariance terms.

The objective of the firm's managers is to find that path of capital stock over time, $K^*(\text{t})$, that maximizes $V_i(0)$. As is fairly well known,[19] the major necessary (first-order) condition is that the optimal path of capital satisfies, at every point in time, the Euler differential equation:

$$(19)\quad \alpha - rk - mA \sum_j A_jK_j \text{ Cov } (u_i, u_j) - [2\beta + 2mA^2 \text{ Var } (u_i)]K - 2r\gamma\dot{K} + 2\gamma\ddot{K} = 0$$

Following Eisner and Strotz's solution procedure, the firm's optimal capital stock is shown in equation (20).[20] Similar to the Eisner-Strotz model, the exponential term approaches zero as t approaches infinity; hence in the limit $K^*(\text{t})$ approaches the stationary value given by the last term in (20). However, our introduction of uncertainty has altered both the stationary or long-run equilibrium value and the speed of adjustment to this equilibrium.

The new long-run equilibrium level of capital differs from the original Eisner-Strotz level by the covariance terms in the numerator and the variance terms in the denominator of the last term of equation (20). As was the case in the examples of the previous section, the introduction of uncertainty has unambiguously lowered the

[19] See the treatment of Eisner and Strotz in their Appendix A, p. 112 ff. More generally, see Richard Courant, Vol. II, ch. 7.

[20] The last term of equation (20) is, of course, the particular solution to the differential equation (19); the exponential term is the part of the homogeneous solution with the negative exponent. Eisner and Strotz argue, p. 80, that the term with the positive exponent drops out because its coefficient must be zero.

$$K^*(t) = \left\{ K(0) - \left[\frac{\alpha - rk - mA \sum_j A_j K_j \operatorname{Cov}(u_i, u_j)}{2\beta + 2mA^2 \operatorname{Var}(u_i)} \right] \right\} \cdot \exp\left[\frac{r - (r^2 + 4\beta/\gamma + 4mA^2 \operatorname{Var}(u_i)/\gamma^2)^{1/2}}{2} \right] t + \frac{\alpha - rk - mA \sum_j A_j K_j \operatorname{Cov}(u_i, u_j)}{2\beta + 2mA^2 \operatorname{Var}(u_i)} \tag{20}$$

level of the long-run equilibrium stock of capital.

In addition to affecting the long-run equilibrium, the introduction of uncertainty has affected the speed of adjustment of the capital stock to this equilibrium. The exponent of e is negative without uncertainty, and it is negative with a larger absolute value with it—the difference being a function of the term $4mA^2 \operatorname{Var}(u_i)/\gamma$. The larger absolute value of the exponent at any time t implies that the actual capital stock is *closer* to the equilibrium value than it would otherwise be. This may sound paradoxical, and it may be; however, two points can be made to make this effect seem plausible. First the *net* effect of an increase in Var (u_j) is to reduce the optimal level of capital, $K^*(t)$, at every point in time; this can easily be verified by finding the derivative $\partial K^*(t)/\partial \operatorname{Var}(u_i)$. Second, the variance term affects $K^*(t)$ exactly as does β, the term related to the (absolute value of the) slope of the demand curve; an increase in this coefficient decreases the long-run equilibrium capital stock and, also, raises the speed of adjustment to the lower equilibrium value—and this was the case for the original Eisner-Strotz model.

III. Conclusions and Recommendations for Further Research

In the first part of this paper I have attempted to transform the Lintner-Mossin-Sharpe (LMS) equation for the value of the neoclassical firm into a form satisfactory for the study, under general conditions, of optimal firm decisions under uncertainty. This involved the consideration of a number of neglected, but important, issues: the determinants of capital gains and the stochastic properties of *future* market parameters such as the riskless rate of interest.

The second part of the paper made a small beginning in applying the revised LMS valuation equation. The equation was expressed in terms of variables controllable by the firm and, for some simple cases, optimal investment and employment policies were explicitly derived.

Although the results presented were largely illustrative, they do suggest, I think, that it will be quite feasible to begin empirical testing for the effects of uncertainty on corporate investment and production decisions. The models developed above are simple ones, but they can incorporate all of the realism of most empirically tested investment models; and they introduce one additional element, mean-variance uncertainty. The one further step that needs to be taken is to develop empirical estimates of the important variances and covariances that might affect firm decisions.

If this general framework is found promising, there are a number of directions in which further research can be pushed.

First, much can be done to sort out which of many potential causes of uncertainty are most important to the firm. Second, there is the fascinating question of the relation of subjective probability distributions to observed outcomes of important variables and to time. Finally, the question of the effect of financial factors on the value of the firm should be faced. In this paper I have assumed models where financial decisions are irrelevant to the maximization of the value of the firm. This does not mean that financial factors cannot or should not be built into a more realistic model of the firm under uncertainty; Tinsley has made a good start on showing just how this might be done.

APPENDIX

FIRST- AND SECOND-ORDER CONDITIONS FOR THE COMPETITIVE FIRM WITH PRICE UNCERTAINTY AND CONSTANT RETURNS TO SCALE

A. *Optimal Levels of Capital and Labor for the Cobb-Douglas Case*

1. We are to maximize equation (13) of the text with respect to K_i and L_i for the case where $Q_i = AK_i^a L_i^{(1-a)}$:

$$(13)\quad E(p_i)Q_i - wL_i - q(-\dot{q}/q + r + d)K_i - m\left[\text{Var}(p_i)Q_i^2 + \sum_{j=1}^{n} Q_iQ_j \text{ Cov}(p_i, p_j)\right]$$

2. Since in the last term Q_j Cov (p_i, p_j) is assumed exogenous for all other firms j, we will abbreviate the last sum as: $Q_i\sum\text{COV}$. Hereafter we will also drop the firm subscripts, i and j.

3. The first-order conditions for the Cobb-Douglas case are:

$$(\text{A1})\quad E(p)aAK^{a-1}L^{1-a} - q(-\dot{q}/q + r + d) - 2m \text{ Var}(p)AK^aL^{1-a}(aAK^{a-1}L^{1-a}) - maAK^{a-1}L^{1-a}\left(\sum \text{COV}\right) = 0$$

$$(\text{A2})\quad E(p)(1-a)AK^aL^{-a} - w - 2m \text{ Var}(p)AK^aL^{1-a}((1-a)AK^aL^{-a}) - m(1-a)AK^aL^{-a}\left(\sum \text{COV}\right) = 0$$

4. Both (A1) and (A2) can be factored into the following form:

$$(\text{A3})\quad \left[E(p) - 2m \text{ Var}(p)Q - m\sum \text{COV}\right] Q_K = q(-\dot{q}/q + r + d)$$

$$(\text{A4})\quad \left[E(p) - 2m \text{ Var}(p)Q - m\sum \text{COV}\right]Q_L = w$$

Thus

$$Q_K/Q_L = \frac{q(-\dot{q}/q + r + d)}{w}$$

5. Since for this linear-homogeneous Cobb-Douglas production function $Q_K/Q_L = (aL/(1-a)K)$, we then have that at any point of optimality:

$$(\text{A5})\quad L^* = \left[\frac{1-a}{a}\right]\frac{q(-\dot{q}/q + r + d)}{w}K^*$$

also shown as (17) in the text.

The labor/capital ratio is independent of the risk variables and the same as the cost minimization ratio under certainty.

6. The above relationship is very useful in deriving the optimal solution (K^*, L^*) for the first-order conditions (A1), (A2). Substituting (A5) for L in equation (A2) we get the following for K^*:

$$\left[E(p) - 2m \cdot \text{Var}(p)AK\left(\frac{q(-\dot{q}/q+r+d)(1-a)}{w\quad a}\right)^{(1-a)} - m\sum \text{COV}\right] \cdot aAK^{a-1}K^{1-a}\left(\frac{q(-\dot{q}/q+r+d)(1-a)}{w\quad a}\right)^{1-a} = q(-\dot{q}/q + r + d)$$

Since $K^{a-1}K^{1-a} = 1$, the above simplifies to the expression for K^* shown as (16) in the text.

B. *The Second-Order Conditions*

1. Letting dV/dK and dV/dL be the first derivatives of (13), the second-order condi-

tions for a maximum are $(d^2V/dK^2)(d^2V/dL^2)-(d^2V/dKdL)^2>0$, and $d^2V/dK^2<0$. Here we shall show that these conditions hold for the general linear homogeneous production function. As is well known, such is not the case for this type of production function under conditions of certainty.

2. We have:

$$\text{(A6)}\quad d^2V/dK^2 = [E(p)-2m\,\text{Var}\,(p)Q-m\textstyle\sum \text{COV}]Q_{KK} - 2m\,\text{Var}\,(p)(Q_K)^2$$

$$\text{(A7)}\quad d^2V/dL^2 = [E(p)-2m\,\text{Var}\,(p)Q-m\textstyle\sum \text{COV}]Q_{LL} - 2m\,\text{Var}\,(p)(Q_L)^2$$

$$\text{(A8)}\quad d^2V/dKdL = [E(p)-2m\,\text{Var}\,(p)Q - m\textstyle\sum \text{COV}]Q_{KL}-2m\,\text{Var}\,(p)Q_KQ_L$$

3. Both (A6) and (A7) are negative, as required, given the expressions for Q_{KK}, Q_{LL}, etc.

4. Let $E(p)-2m\,\text{Var}\,(p)Q-m\sum \text{COV}$ be represented by [ ]. Then the expression $(d^2V/dK^2)(d^2V/dL^2)-(d^2V/dKdL)^2$ equals:

$$\begin{aligned}&[\ ]^2Q_{KK}Q_{LL}-2m\,\text{Var}\,(p)[\ ]Q_{LL}(Q_K)^2\\ &\quad -2m\,\text{Var}\,(p)[\ ]Q_{KK}(Q_L)^2\\ &\quad +4m^2(\text{Var}\,(p))^2(Q_L)^2(Q_K)^2\\ &\quad -[\ ]^2(Q_{KL})^2+4m\,\text{Var}\,(p)[\ ]Q_KQ_LQ_{KL}\\ &\quad -4m^2(\text{Var}\,(p))^2(Q_K)^2(Q_L)^2\end{aligned}$$

5. The fourth and seventh terms cancel out, and since for the linear homogeneous production function $Q_{KK}Q_{LL}-(Q_{KL})^2=0$, the first and fifth terms equal zero. Hence the above second-order conditions reduce to:

$$-2m\,\text{Var}\,(p)[\ ]Q_{LL}(Q_K)^2 - 2m\,\text{Var}\,(p)[\ ]Q_{KK}(Q_L)^2 + 4m\,\text{Var}\,(p)[\ ]Q_KQ_LQ_{KL}$$

6. For the first-order conditions to be satisfied [ ] must be greater than zero. Var(p) is greater than zero and m is greater than zero. So the sign of the second-order condition depends on the sign of:

$$4Q_KQ_LQ_{KL} - 2(Q_K)^2Q_{LL} - 2(Q_L)^2Q_{KK}$$

7. All three terms are greater than zero for any linear-homogeneous production function. So the second-order conditions for a maximum are satisfied.

REFERENCES

C. Bischoff, "Business Investment in the 1970's: A Comparison of Models," *Brookings Papers*, Washington 1971, *1*, 13–64.

R. Courant, *Differential and Integral Calculus*, Vol. II, New York 1961.

R. Eisner and R. H. Strotz, "Determinants of Business Investment," in *Impacts of Monetary Policy, Commission on Money and Credit*, Englewood Cliffs 1963.

E. Fama, "Multiperiod Consumption-Investment Decisions," *Amer. Econ. Rev.*, Mar. 1970, *60*, 163–74.

———, "Risk, Return and Equilibrium," *J. Polit. Econ.*, Jan./Feb. 1971, *79*, 30–55.

R. Hamada, "Portfolio Analysis, Market Equilibrium, and Corporation Finance," *J. Finance*, Mar. 1969, *24*, 13–31.

J. Hirshleifer, *Investment, Interest and Capital*, Englewood Cliffs 1970.

M. Jensen and J. Long, Jr., "Corporate Investment under Uncertainty and Pareto Optimality in the Capital Markets," *Bell J. Econ.*, Spring 1972, *3*, 151–74.

D. Jorgenson and C. Siebert, "Optimal Capital Accumulation and Corporate Investment Behavior," *J. Polit. Econ.*, Nov./Dec. 1968, *76*, 1123–51.

J. Lintner, (1965a) "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Rev. Econ. Statist.*, Feb. 1965, *47*, 13–37.

———, (1965b) "Security Prices, Risk and Maximal Gains from Diversification," *J. Finance*, Dec. 1965, *20*, 587–615.

———, "Optimum or Maximum Corporate Growth under Uncertainty," in R. Marris and A. Wood, eds., *The Corporate Economy: Growth, Competition and Innovative Potential*, Cambridge, Mass. 1971.

R. Litzenberger and A. P. Budd, "Corporate Investment Criteria and the Valuation of

Risk Assets," *J. Finance Quant. Anal.*, Dec. 1970, *5*, 395–419.

M. Miller and F. Modigliani, "Dividend Policy, Growth and the Valuation of Shares," *J. Bus., Univ. Chicago*, Oct. 1961, *34*, 411–33.

F. Modigliani and M. Miller, "The Cost of Capital, Corporation Finance and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261–97.

J. Mossin, "Equilibrium in a Capital Asset Market," *Econometrica*, Oct. 1966, *34*, 768–83.

M. Rothschild and J. Stiglitz, "Increasing Risk: II. Its Economic Consequences," *J. Econ. Theory*, Mar. 1971, *3*, 66–84.

A. Sandmo, "On the Theory of the Competitive Firm Under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65–73.

W. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *J. Finance*, Sept. 1964, *19*, 425–42.

G. Stevens, "On the Value of the Firm, Discounting, and Optimal Investment Under Uncertainty," spec. stud. pap. 26, rev., Fed. Res. Board, Washington 1973.

J. Stiglitz, "On the Optimality of the Stock Marget Allocation of Investment," *Quart. J. Econ.*, Feb. 1972, *86*, 25–60.

H. Theil, "A Note on Certainty Equivalence in Dynamic Planning," *Econometrica*, Apr. 1957, *25*, 347–49.

P. Tinsley, "Capital Structure, Precautionary Balances, and the Valuation of the Firm: The Problem of Financial Risk," *J. Finance Quant. Anal.*, Mar. 1970, *5*, 33–62.

E. Zabel, "Efficient Accumulation of Capital for the Firm," *Econometrica*, Jan.-Apr. 1963, *31*, 131–50.

Land Use and Zoning in an Urban Economy

By WILLIAM J. STULL*

Zoning statutes are a nearly ubiquitous phenomenon in urban America. Allen Manvel in his report to the National Commission on Urban Problems, p. 31, estimated that 90 percent of the cities and towns in the United States with a population of over 5,000 had some kind of zoning statute in 1967. In addition, 51 of the 52 largest cities possessed such ordinances.[1] In spite of this ubiquity, and in spite of the recent public controversy concerning the legality of suburban exclusionary zoning, the modern economics literature on the institution is amazingly sparse. There have been some significant contributions in recent years, notably those by Otto Davis, Martin Bailey, Davis and Andrew Whinston, and John Crecine, Davis, and Robert Jackson. However, I think it is fair to say that economists have barely scratched the surface of what seems to me to be a very complex subject. In particular, there has been very little progress made in the task of incorporating zoning or land use parameters into the kinds of micro-economic models that economists are accustomed to deal with. The contrast with another urban economic institution—the property tax—is rather striking in this respect.

The basic objective of this paper will be to attempt to fill this gap. The first section introduces a simple model of an urban space economy with a manufacturing and residential sector. Having constructed this model, I will show how a change in the allocation of land between the two sectors affects the spatial pattern of land rents, the size of the urban labor force, and the wage rate. The second section of the paper will then be devoted to a discussion of the appropriateness and implications of a land allocation policy which maximizes aggregate land rent.

I. The Basic Model

Let us begin by assuming that a city has emerged on a line which extends indefinitely in two directions. Except for a fixed center point, which represents a transportation node connecting our urban economy with the outside world, the line is featureless.[2] Following William Alonso, p. 18, we shall also assume that all of the city's economic transactions, both in the labor market and in commodity markets, take place at this center point. Land in this economy has only two potential economic uses: as a location for residential activities or as a location for manufacturing activities. The spatial structure of the city is under the control of a private profit-maximizing developer who lays it out and then each year rents at an auction the available parcels of land to households and firms who wish to locate there. The equilibrium state of the economy for a given

---

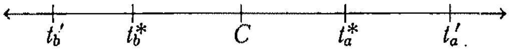
* Department of economics, Swarthmore College. I am indebted to Jerome Rothenberg, John Harris, Robert Solow, Ronald Grieson, Larry Hirschhorn, William Wolfson, George Borts, and an anonymous referee for useful comments on earlier drafts. I would also like to thank Bruce Char for editorial assistance.

[1] See Manvel, p. 171. The one exception is Houston. For a discussion of the Houston case, see Bernard Siegan.

[2] This is a standard assumption. See, for example, William Alonso, pp. 15–17. As used here and by Alonso, "featureless" means the permanent absence of both distinctive natural features (topographical, geological, and climatic) and all man-made improvements (including structures).

year is thus jointly determined by the layout imposed by the developer for that year and the outcome of the annual land auction.

To simplify the subsequent analysis, we shall suppose that most of the actual laying out of the town takes place prior to any auctions. There are two steps to this preliminary subdivision. First, the developer divides the entire line into identical lots of length q^*. Second, he groups these parcels into homogeneous use districts. Manufacturing firms are constrained to locate near the urban center, while households are constrained to outlying locations. An example of the resulting land use pattern is exhibited below.



C is the center of the city. Manufacturing is restricted to the interval (t_b^*, t_a^*), while residences are restricted to the open-ended segments extending to the left from t_b^* and to the right from t_a^*. The natural operation of the market for residential land then produces two "housing" zones: (t_b', t_b^*) and (t_a^*, t_a'). All parcels to the right of t_a' and to the left of t_b' remain vacant. Associated with this partition there is a market rent gradient (not shown) indicating the equilibrium rent paid to the developer at each location. Note that the developer may adjust the exact location of t_a^* and t_b^* from one auction to another (and thereby effect changes in the rent gradient and t_a' and t_b'). We will assume, however, that he does not change either the lot sizes or the homogeneity and relative location of the two types of districts once these have been initially determined.

In addition to a land market this model possesses a competitive labor market. Labor is demanded by the manufacturing sector in accordance with the world price of the output which it produces, the equilibrium urban wage rate, and the amount of land allocated to manufacturing. Labor is supplied by the residential sector in accordance with the wage rate and the extent to which "desirable" residential locations are available in the economy. Fluctuations in the quantity of labor supplied are ultimately brought about by migration into or out of the urban area in response to changed conditions there or in the rest of the world.

The way in which land use problems enter the model may now be easily seen. Prior to any auction, the developer must choose the appropriate locations for t_a^* and t_b^*, the manufacturing-residential boundary points. His choice will be conditioned by the fact that there is a dual relationship between these two sectors. The first is through the labor market; the second comes about because, other things equal, consumers do not like to live in communities which have extensive manufacturing activities in them.[3] The developer must then weigh these two factors and choose a pair of boundary points which maximizes his welfare function. A similar kind of tradeoff faces local governments when they try to decide how to zone available vacant land in their communities. Before we can discuss such decision problems intelligently, however, it will be necessary to go over the details of the basic model rather carefully. This task will occupy us for the remainder of Section I.

A. *Residential Land and Labor Supply*

All consumers are assumed to have

---

[3] This is a neighborhood externality effect of the sort discussed in the urban renewal literature. See, for example, Jerome Rothenberg or Davis and Whinston (1961). There is currently a controversy over the empirical significance of these effects. Both John Kain and John Quigley and I (1972, ch. 4) present evidence to support the view that neighborhood externalities exert a significant influence on land values. Crecine, Davis, and Jackson come to the opposite conclusion. In this paper, it will be assumed that environmental amenities do in fact affect both household locational choices and land rent.

identical utility functions of the following form:

$$u = u(x_1, x_2, \ldots, x_n, q, t, t - t^*) \tag{1}$$

where x_i = the quantity of ordinary commodity i purchased
q = the quantity of land rented
t = the distance from the consumer's residence to the urban center C
t^* = the distance from the nearest zoning boundary[4] to C

The inclusion of x_i and q is quite standard and requires no commentary. The variable t is included in (1) as a way of indirectly capturing the loss of leisure brought about by the need to devote some time to shopping and work trips (all of whose destinations will be at C). The last argument in (1) formally introduces neighborhood externalities into the model. Other things equal, households prefer living far away from industrial zones to living near them. The assumed signs of the first-order partial derivatives are the natural ones:

$$\begin{aligned} &u_i > 0 \qquad (i = 1, 2, \ldots, n) \\ &u_{n+1} > 0 \\ &u_{n+2} < 0 \\ &u_{n+3} > 0 \end{aligned} \tag{2}$$

Each consumer is also subject to a budget restraint:

$$y = \sum p_i x_i + mq + h(t) \tag{3}$$

where y = the annual urban money wage rate
p_i = the price of ordinary commodity i
m = the annual land rent payment per unit land
h = the transportation cost function $[h'(\ ) > 0]$

Each household's income is thus exhausted by payments for ordinary commodities, for residential land, and for transportation. This specification assumes that the total annual expenditure for the last of these is a function only of the distance from the household's residence to the center of the city. This follows from our previous assumption that all economic transactions, including both household purchases of ordinary commodities and sales of labor services, take place at C.

The actual assignment of households to lots takes place at a land auction. Recall first that the size of the lot has been fixed in advance. Thus,

$$q = q^* \tag{4}$$

a constant at all locations. When the auction is announced, consumers migrate to the city from the rest of the world and bid on the parcels available. We shall suppose that the new city is small so that it is reasonable to assume that there is an infinitely elastic supply of prospective residents at some particular utility level, say $\bar{u}$. This assumption, coupled with (4), implies the following equilibrium condition for each household once the auction is complete:

$$\bar{u} = u(x_1, x_2, \ldots, x_n, q^*, t, t - t^*) \tag{5}$$

Thus we are assuming that migration equalizes utility levels across cities. This in turn is a somewhat more general statement of the familiar proposition that migration equalizes real wage rates.

We now have three functional restrictions on the equilibrium behavior of any household in our economy. First, it must always be on its budget line (3). Second, the quantity of land purchased is fixed in advance and constant at all locations (4). And third, it must always be at a fixed utility level (5). These three conditions, together with the assumption that households are rational in their bidding behavior, are enough to determine a gradient showing how the equilibrium rent per unit of residential land varies with distance. We will assume that the developer always lays out a symmetrical city. Consequently, it is only necessary here to focus on one side of this overall relationship. Let us designate the function for the right-hand rent gradient as follows:[5]

---

[4] We shall also from time to time use t^* as a symbol for the boundary point itself.

$$m = m(t; y, t^*) \tag{6}$$

A useful way to think about this relationship is as a curve in rent-distance space which shifts up or down in response to changes in y or t^*.

To see how this function is obtained let us choose some particular residential location (thus fixing t) and assume for the moment that y, t^*, $\bar{u}$, p_i $(i=1, 2, \ldots, n)$, and $h(\ )$ are given. Next, substitute (4) into (3) and solve for m to obtain

$$m = [1/q^*][y - \sum p_i x_i - h(t)] \tag{7}$$

Consumers are assumed to be rational; consequently, the equilibrium rent paid at the location in question will be the maximum rent a consumer can pay and still be at utility level $\bar{u}$. This suggests that the following Lagrangian function be set up:

$$\begin{aligned} R = {} & [1/q^*][y - \sum p_i x_i - h(t)] \\ & - \lambda[\bar{u} - u(x_1, \ldots, x_n, q^*, t, t-t^*)] \end{aligned} \tag{8}$$

Since t was fixed a priori, the only decision variables under the control of the household are the x_i. Consequently, the first-order conditions for a maximum are

$$\begin{aligned} \partial R/\partial x_i &= [-p_i/q^*] + \lambda u_i = 0 \qquad (i=1, 2, \ldots, n) \\ \partial R/\partial \lambda &= \bar{u} - u(x_1, \ldots, x_n, q^*, t, t-t^*) \\ &= 0 \end{aligned} \tag{9}$$

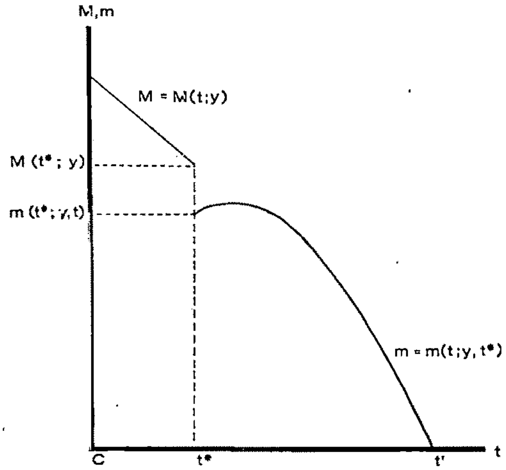


FIGURE 1. MANUFACTURING AND RESIDENTIAL RENT GRADIENTS

These equations form a system of $n+1$ equations which may be solved to obtain values for the $n+1$ unknowns, λ and the x_i. The equilibrium rent per unit land may then be found by substituting the values of the x_i into (7).

The signs of the partial derivatives of (6) may be determined by making use of the envelope theorem. In the present context this theorem states that

$$\partial m/\partial \alpha = \partial R/\partial \alpha \tag{10}$$

for any parameter α. Using it, let us first of all obtain an expression for the slope of the land rent gradient.

$$\begin{aligned} \partial m/\partial t = \partial R/\partial t = {} & [1/q^*][-h'(t)] \\ & + \lambda[u_{n+2} + u_{n+3}] \end{aligned} \tag{11}$$

The sign of (11) is indeterminate because of the externality effect. If households did not care how far they lived from factories, then u_{n+3} would be equal to zero and the conventional result, a downward sloping rent gradient, would be obtained.[6]

In Figure 1 an arbitrary residential rent gradient for one-half of the city is depicted.

---

[5] Note that the parameters which play no important role in the subsequent analysis (such as $\bar{u}$ and p_i) have been incorporated into the functional form of $m(\ )$ rather than being directly included in the argument list. It turns out that $m(\ )$ is mathematically equivalent to Alonso's residential bid price curve. See his ch. 4 and my 1973 paper.

[6] The sign of λ is known to be positive from (9).

The curve is drawn so that it slopes upward in that section of the residential zone located close to the manufacturing sector. There is no necessity about this particular shape, however. Indeed, a rather likely real world situation is one in which transport costs are so high that they swamp the externality effects in (11), causing the gradient to be negatively sloped everywhere.

Next, let us consider how the rent of an arbitrary residential parcel would change if the developer decided to move the manufacturing-residential boundaries outward. Using the same procedure as before we obtain

$$\partial m/\partial t^* = -\lambda u_{n+3} < 0 \tag{12}$$

This result tells us that the larger the developer makes the manufacturing district prior to an auction, the lower will be the rent he receives for each residential parcel (other things equal, including the wage rate). This comes about simply because of the externality effect $u_{n+3}>0$ in equation (2). In terms of Figure 1, an increase in t^* causes the residential rent gradient to shift *downward* at every point. This in turn causes t', the urban-rural margin, to shift to the left. Thus we have also shown (informally) that

$$\partial t'/\partial t^* < 0 \tag{13}$$

Recall that thus far we have assumed that the wage rate received by all households in our economy is constant. If we now allow a small *ceteris paribus* increase in this variable, how will the land rent gradient shift? Proceeding as above,

$$\partial m/\partial y = 1/q^* > 0 \tag{14}$$

Basically, (14) tells us that income increases are completely capitalized into land rent, a result which is a natural consequence of our earlier assumptions about intercity migration and the equalization of utility levels. Once again referring to Figure 1, it should be clear that an increase in y causes the entire land rent gradient to shift *upward*. It follows immediately, using an argument identical to the one underlying (13), that

$$\partial t'/\partial y > 0 \tag{15}$$

We are now in a position to say something about the supply curve of labor in this economy. Given t^* and y (and the other parameters), the labor force of the city is

$$L_s = 2(t' - t^*)/q^* = L_s(y; t^*) \tag{16}$$

assuming a symmetrical city and one worker per household. Applying (15), we obtain

$$\partial L_s/\partial y = (2/q^*)(\partial t'/\partial y) > 0 \tag{17}$$

This result tells us that the labor supply curve in our economy is *positively sloped.* In terms of Figure 1, an increase in y causes $m(\ )$ to shift upward, thereby increasing the quantity of inhabited land $(t'-t^*)$. Since population density is constant, this necessarily implies an increase in the labor force.

B. *Manufacturing Land and Labor Demand*

Now let us turn to the manufacturing sector of our urban economy. We assume first that all firms are perfectively competitive and have identical production functions. In addition, as indicated earlier, land in the manufacturing zone has been subdivided in advance into identical lots of size q^*, each one of which is auctioned off to a different firm. Thus,

$$Z = F(L, q^*) \tag{18}$$

where Z = the quantity of output produced
L = the quantity of labor input.

The derivatives of $F(\ )$ have the usual signs:

$$F_1 > 0 \qquad F_{11} < 0 \tag{19}$$

Each firm hires labor locally at the current urban wage rate and sells its output at the world price. Before a firm's output can be sold, however, it must be shipped from the point of manufacture to C, the interurban transportation node. Because these local transportation costs are absorbed by the seller, a site close to C is, other things equal (including land rent), more desirable than one further out. Annual profit under these conditions is defined as follows:

$$P = [W - k(t)]Z - yL - Mq^* \tag{20}$$

where W = the world price of the output
k = the transportation cost function $[k'(\ ) > 0]$
t = the distance from the point of production to C
M = the annual land rent payment per unit land.

At the land auction each parcel in the manufacturing sector is rented to the highest bidder. Since we shall assume that our new town faces an infinitely elastic supply of firms willing to earn a "normal" zero profit, the following relationship holds at every location:

$$M = [W - k(t)][Z/q^*] - y[L/q^*] \tag{21}$$

To obtain this result, (20) was rearranged and P was set equal to zero. Finally, we shall assume that all firms are bid-maximizers. Thus, in equilibrium each firm will pay the maximum rent possible consistent with the earning of a zero profit. Since the quantity of labor hired is the only decision variable in the problem, there is only one first-order condition applicable to each firm:

$$\partial M/\partial L = [W - k(t)][F_1/q^*] - [y/q^*] = 0 \tag{22}$$

Equations (18), (21), and (22) constitute a system of three equations in three unknowns, M, L, and Z. If this system is solved, two equations can be obtained showing how M and L separately vary with parameters t and y. Denote the first of these relationships:[7]

$$M = M(t, y) \tag{23}$$

A routine calculation reveals that

$$\begin{aligned} \partial M/\partial t &= [-k'(t)/q^*]F(L, q^*) < 0 \\ \partial M/\partial y &= [-L/q^*] < 0 \end{aligned} \tag{24}$$

Thus, the rent gradient is everywhere negatively sloped in the manufacturing zone (as shown in Figure 1). In addition, a *ceteris paribus* increase in the urban wage rate causes this gradient to shift downward at every point.

We denote the second relationship in an analogous way.

$$L = L(t; y) \tag{25}$$

Repeating the same calculation, we obtain

$$\begin{aligned} \partial L/\partial t &= [k'(t)/(W - k(t))][F_1/F_{11}] < 0 \\ \partial L/\partial y &= [1/F_{11}][1/(W - k(t))] < 0 \end{aligned} \tag{26}$$

As one might anticipate, the labor gradient declines with distance from C; accessible parcels have more labor employed on them than do inaccessible parcels. In addition, this gradient shifts downward at every point with a *ceteris paribus* increase in y. These conclusions follow from (19) and the recognition that the cost of transporting one unit of output from the firm's location to C will never exceed the price this unit receives once it reaches its destination.

Let us now consider the demand curve for labor in our economy. The total quantity of labor demanded by the manufacturing sector at some wage rate y and for some particular boundary point t^* is

$$L_d = (2/q^*) \int_0^{t^*} L(t; y)\,dt = L_d(y; t^*) \tag{27}$$

---

[7] Both (23) and (25) represent gradients for the right-hand side of the city only. Again, parameters which play no role in the subsequent analysis have been incorporated into the functions themselves.

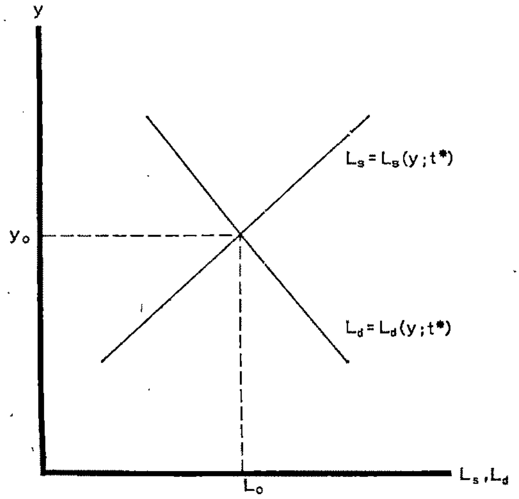


FIGURE 2. EQUILIBRIUM IN THE LABOR MARKET

It follows immediately from $\partial L/\partial y$ in (26) that:[8]

$$(28)\quad \partial L_d/\partial y = (2/q^*)\int_0^{t^*}[\partial L(t;y)/\partial y]dt < 0$$

Thus the demand curve for labor in our economy is negatively sloped.

Both the supply and demand curves for labor are shown superimposed in Figure 2. The equilibrium wage rate and labor force (y_0, L_0) are determined by their intersection. The impact of any change in an exogenous variable on the urban labor market may then be analyzed by determining its effect on the locations of these two curves.

C. *Boundary Adjustments*

Since this paper is ultimately concerned with land use policies, a natural question to ask is how the equilibria in the labor and land markets will be affected by a change in t^*, the manufacturing-residential boundary point. Let us start by assuming that Figures 1 and 2 show the land and labor market equilibria achieved as a result of last year's auction. Let us assume further that our developer has decided that the particular t^* which he chose last year was smaller than the t^* which would have maximized his profits. Consequently, prior to this year's auction he moves the two manufacturing-residential boundary points out to what he believes to be the optimal locations. Assuming that all other conditions remain the same, what will be the impact of this change?

Consider the labor market first. On the supply side we have from (16) and (13)

$$(29)\quad \partial L_s/\partial t^* = [2/q^*][(\partial t'/\partial t^*) - 1] < 0$$

Thus a small increase in t^* causes the supply curve of labor to shift upward and to the left. The logic behind this result is straightforward. Basically, an increase in t^* (other things equal) makes our new town a less desirable place to live for the typical citizen. Consequently, at each wage rate fewer people will wish to migrate to the city than would have if the change had not occurred. This decrease in desirability comes about for two reasons. First, the parcels of residential land which formerly were most proximate to the workplace—i.e., those adjoining t^*—are no longer available to households because they have been converted to manufacturing uses. Second, the externality effect described earlier increases with an increase in t^*, thereby causing all sites remaining in residential use to become less attractive than previously. The combined effect can easily be seen by examining Figure 1. Increasing the amount of land devoted to manufacturing simultaneously moves t^* to the right and t' to the left, thus effecting a shrinkage in both the quantity of inhabited residential land and the labor force.

The demand for labor is given in (27). It follows immediately that

---

[8] Here and elsewhere in this paper, I make use of the following relationship:

$$d/da\int_p^q f(x,a)dx = \int_p^q \partial/\partial a[f(x,a)]dx + f(q,a)(dq/da) - f(p,a)(dp/d)a$$

$$\partial L_d/\partial t^* = (2/q^*)L(t^*; y) > 0 \tag{30}$$

Thus, an increase in t^* causes the demand curve for labor to shift upward and to the right. At each wage rate, more workers than before are demanded by the manufacturing sector. This is a sensible result because increasing t^* allows more firms to participate in the economy.

It should now be clear how a small increase in t^* will affect the city's wage rate and labor force. Both the demand and supply curves for labor in Figure 2 will shift upward. Thus our developer will find that allocating more land to manufacturing this year will cause this year's equilibrium money wage to be greater than last year's. That is,

$$dy/dt^* > 0 \tag{31}$$

The effect of the change on the size of the equilibrium labor force and population, on the other hand, will be indeterminate. As any introductory student knows, a simultaneous upward shift in both supply and demand has a positive effect on price but an ambiguous effect on quantity. Total differentiation of the supply and demand equations (not shown here) reveals that the labor force will tend to increase if labor supply is elastic with respect to the wage rate and labor demand is inelastic. In addition, it will also tend to increase if the externality effect is weak and if the additional firms entering after the increase in t^* demand large amounts of labor. More precise conclusions would require knowledge of the specific functional forms involved.

The results in the previous paragraph have some interesting implications. First of all, we have demonstrated that there is a connection between the quantities of land allocated to different activities and certain other important economic variables. This suggests that zoning and other land policies should not be implemented in isolation but rather should be coordinated with existing economic planning or development programs being carried out in other governmental departments. This is hardly a new idea as far as small communities are concerned. Very often they will regard the local zoning ordinance as the only effective way of controlling economic activity within their boundaries. Large cities, on the other hand, typically have separate zoning and economic development (or redevelopment) agencies which operate quite independently of each other.[9] Our analysis suggests that this is not a sensible way to run a city.

Secondly, note that our city can be in equilibrium with the rest of the world at any of an infinity of different money wage rates (each associated with a different land use policy). This result is a natural consequence of our assumption that it is utility levels rather than wage rates which are equalized between cities by interurban migration. It also suggests why it is necessary for social scientists studying interurban migration to deflate money wage rates with amenity and price level indices.

The next step is to trace the labor market changes described above back through to the underlying land market adjustments. The complete effect of the increase in t^* on the rent gradients in the two zones of the economy is complex. Because of (24) and (31) an increase in t^* indirectly causes the equilibrium rent for each parcel in the manufacturing zone to decline; that is,

$$dM/dt^* = (\partial M/\partial y)(dy/dt^*) < 0 \tag{32}$$

The effect on the residential land rent gradient, however, is indeterminate. Differentiating, we obtain

$$dm/dt^* = (\partial m/\partial y)(dy/dt^*) + \partial m/\partial t^* \tag{33}$$

The first term of (33) is positive ((14) and (31)), but the second is negative (12).

The logic of these results is straightforward-ward. As shown above, an increase in t^* causes the wage rate to rise. This makes manufacturing sites in the city less desirable than before; consequently, rents in the manufacturing sector will fall. Rents in the residential zone, however, may either decrease or increase depending upon whether or not the (negative) externality effect outweighs the (positive) induced income effect. A likely outcome is one in which rents in different parts of the residential zone move in different directions.

---

[9] For some evidence supporting this assertion in the case of Boston, see William Weismantel.

Finally, we may ask what difference the neighborhood externality assumption makes to all of this analysis. The basic fact is that the existence of consumer tastes about the location of manufacturing land causes the size of the urban labor force to be more sensitive to land use changes than would otherwise be the case. As is clear from (29), if there were no neighborhood externalities, a marginal increase in t^* will induce a *smaller* upward shift in the labor supply curve than would occur if externalities did exist. This result suggests in turn that, other things equal, cities in the real world which are characterized by particularly offensive man-made environments (due to pollution, ugliness, etc.) ought to have higher prevailing money wage rates and smaller labor forces in equilibrium than those cities which are not so characterized.

II. Maximizing Aggregate Land Rent

A. *The Developer Case*

In the preceding section, we concerned ourselves with the structural characteristics of a simple model of an urban economy. Nothing was said, however, about the behavior of the developer, who in fact was in charge of planning the town and who in the end had to make all land allocation decisions. Where would he choose to locate the manufacturing-residential boundary points? Presumably our developer wishes ultimately to maximize annual profits. However, once the costs of purchasing and subdividing the land are sunk, a strategy which maximizes annual profits is equivalent to one which maximizes aggregate land rent. Thus we will assume that our developer chooses the t^* which maximizes V, where

$$(34)\quad V = \int_0^{t^*} M(t; y)dt + \int_{t^*}^{t'} m(t; y, t^*)dt$$

The first term gives the total rental receipts in the manufacturing zone, the second the total rental receipts in the residential zone (both for the right-hand side of the city only).

A necessary condition for the maximization of V is that its first derivative with respect to t^* be equal to zero. A straightforward, but tedious, calculation reveals that[10]

$$(35)\quad dV/dt^* = M(t^*; y) - m(t^*; y, t^*) + \int_{t^*}^{t'} \partial m/\partial t^* dt$$

Consequently, at the optimal t^*

$$(36)\quad M(t^*; y) = m(t^*; y, t^*) - \int_{t^*}^{t'} \partial m/\partial t^* dt$$

Essentially, (36) is a restriction on the rents per unit land paid by the two uses at the boundary point. As shown in Figure 1, $M(t^*; y)$ is the rent paid by the manufacturing firm located immediately to the left of t^*, and $m(t^*; y, t^*)$ is the rent paid by the corresponding household immediately to the right of t^*.

Note, first of all, that if there are no externalities in the economy the integral term on the right-hand side of (36) is equal to zero. Consequently, a necessary condition for the maximization of aggregate land rent in the community is that the rents paid by the two uses at the boundary point be equal. If there are externalities in the economy, however, the integral term in (36) is negative. In this case a necessary condition for a maximum is that the manufacturing rent at t^* be greater than the residential rent at the same point, with the difference being equal to the total decrease in residential land rents which would come about if the boundary point were moved out slightly and y were held constant. This last decrease is of course the consequence of neighborhood effects. The reader will recall from the previous section how an outward shift in t^*, other things equal, caused the entire residential rent gradient to shift downward. In either case it is now clear that the developer will determine the optimal allocation of urban land between the two uses by solving (36) for t^*.

---

[10] This result requires repeated application of the formula given in fn. 8. A copy of the calculations may be obtained from the author. For a discussion of a similar result in a residential segregation context, see Bailey.

B. *The Community Case*

Suppose we now drop the assumption that all land in the city is owned and controlled by a private developer and instead assume decentralized ownership and community regulation. Is it reasonable under these circumstances to assume that the community will behave like the developer described above and choose the manufacturing-residential boundary point which maximizes aggregate land rent? This is in fact a complex question and one whose full exploration is beyond the scope of this paper. Nonetheless, the model developed earlier can be used to shed light on some of the important issues which must be faced in any attempt to answer it.

The crucial consideration is the extent to which the returns from a rent maximizing strategy will accrue to the typical community resident. There are a number of different possibilities here. For one thing, the typical resident may be a tenant rather than a landlord. As is well known, tenant interest in land use policy is ordinarily rather limited. There are two reasons for this. First, tenants are usually transients and, second, even when they are not, periodic renegotiation of rental contracts means that all locational advantages and disadvantages brought about by a zoning change will eventually be fully reflected in the rent paid, leaving each tenant no better off than before the change. Consequently, in a community occupied wholly by tenants there is no particular reason why a strategy which maximizes aggregate land rent would be preferred to any other strategy—unless it also happened to maximize side payments from landlords to tenants.

Further, even if all residents of the community were local landlords, the maximization of aggregate land rent (defined to include implicit rent in the case of owner occupiers) is not necessarily the most desirable strategy. For one thing, in this situation there still may be significant quantities of land held by absentee landlords. Consequently, some of the fruits of a rent maximization strategy will accrue to outsiders in the form of capital gains. Unless mechanisms exist for the capture of these gains (for example, a property tax law which permitted the nonresidential tax rate to be greater than the residential rate), a community might rationally prefer maximizing the aggregate rental value of local property held by residents to maximizing the aggregate rental value of all property. Secondly, even if all land in the community is also owned by local residents, a rent maximizing strategy will not be desirable if its intracommunity redistributive consequences are adverse. This is clearly a possibility given the complex response of the community rent structure to boundary changes.

The upshot thus is that there are many situations where it might not be plausible to assume that a community will pursue the land use policy which maximizes aggregate land rent.[11] Nonetheless, such an assumption is not patently unreasonable in all cases and several authors have seen fit to propose it, though with varying degrees of ethical approval.[12] If this strategy can be accepted, the solution worked out in the previous subsection applies. If not, then alternative optimality criteria must be specified and their consequences worked out using a model such as the one developed in the first section of this paper.

A final observation is that even if it is not reasonable to assume that a community will pursue a rent maximizing strategy, it is still highly likely that it will wish to impose locational restrictions on those land uses which generate strong negative neighborhood effects. These restrictions are likely, in turn, to lead to price discontinuities of the type shown in Figure 1. As has been suggested by Bailey, such discontinuities pose enforcement problems in the community case which do not exist in the developer case. In particular, they create strong incentives for the landlords immediately to the right of t^* to try to get their land rezoned so that they may achieve capital gains through conversion. Thus when land ownership is decentralized, our model predicts that the existence of adverse neighborhood effects will create a conflict between public and private interests in the maintenance of any particular zoning boundary. Such conflicts can, in turn, account for the widespread corruption of local zoning officials which many Americans seem to take for granted.[13]

---

[11] James Ohls, Richard Weisberg, and Michelle White reach a similar conclusion in a paper which explores the rent maximization hypothesis in some depth.

[12] For example, one of the two zoning "theories" (i.e., principles for guiding local zoning policy) discussed by Richard Babcock is community land value maximization. Also Bailey's classic paper implicitly endorses this strategy on efficiency grounds.

[13] John Delafons discusses this phenomenon in his well-known survey of American land use control practices, pp. 8–9. A representative exposé may be found in Alfred Balk.

REFERENCES

W. Alonso, *Location and Land Use*, Cambridge 1964.

R. Babcock, *The Zoning Game*, Madison 1966.

A. Balk, "Invitation to Bribery," *Harper's*, Oct. 1966, *233*, 18–24.

M. Bailey, "Note on the Economics of Residential Zoning and Urban Renewal," *Land Econ.*, Aug. 1959, *33*, 288–92.

J. Crecine, O. Davis, and R. Jackson, "Urban Property Markets: Some Empirical Results and Their Implications for Municipal Zoning," *J. Law Econ.*, Oct. 1967, *10*, 79–99.

O. Davis, "Economic Elements in Municipal Zoning Decisions," *Land Econ.*, Nov. 1963, *39*, 375–86.

——— and A. Whinston, "The Economics of Complex Systems: The Case of Municipal Zoning," *Kyklos*, 1964, *17*, Fasc. 3, 419–46.

——— and ———, "The Economics of Urban Renewal," *Law Contemp. Prob.*, Winter 1961, *26*, 105–17.

J. Delafons, *Land-Use Controls in America*, Cambridge 1962.

J. Kain and J. Quigley, "Measuring the Value of Housing Quality," *J. Amer. Statist. Assn.*, June 1970, *65*, 532–48.

A. Manvel, *Local Land and Building Regulations*, research report no. 6, Nat. Comm. Urban Problems, Washington 1967.

J. Ohls, R. Weisberg, and M. White, "The Effect of Zoning on Land Value," mimeo., Princeton Univ. 1973.

J. Rothenberg, *Economic Evaluation of Urban Renewal*, Washington 1967.

B. Siegan, "Non-Zoning in Houston," *J. Law Econ.*, Apr. 1970, *13*, 71–147.

W. Stull, "An Essay on Externalities, Property Values, and Urban Zoning," unpublished doctoral dissertation, Mass. Inst. Technology 1972.

———, "A Note on Residential Bid Price Curves," *J. Reg. Sci.*, Apr. 1973, *13*, 107–13.

W. Weismantel, "Collision of Urban Renewal with Zoning: The Boston Experience, 1950–1967," unpublished doctoral dissertation, Harvard Univ. 1968.

The Neoclassical Production Function

By LOWELL GALLAWAY AND VISHWA SHUKLA*

The concept of the neoclassical production function has come under severe attack in recent years, particularly from the so-called Cambridge critics of neoclassical theory.[1] The essence of the argument can be illustrated by reference to Figure 1 which depicts hypothetical relationships between wage and interest (or profit) rates for different systems of production. The standard neoclassical conceptualization holds that at successively higher interest rates there is a "switch" to techniques of production that economize on capital inputs.[2] This simple formulation has been attacked on two grounds. First, it has been argued (and illustrated by appropriate counterexamples) that it is possible for a single production technique to be the most profitable at two different points on the production frontier that is generated by the full set of available production techniques.[3] This is the well-known reswitching argument, the acceptance of which destroys the notion of a well-ordered production function in which there is a systematic relationship between factor prices and the capital-labor ratio. The second line of attack has raised the possibility that "price" Wicksell effects will dominate "real" Wicksell effects and destroy the well-ordered character of the production function. G. C. Harcourt has demonstrated that this will be the case if the relationship between the wage and interest rates shown in Figure 1 is concave with respect to the origin. On the other hand, convexity of this relationship implies dominance of real Wicksell effects.

The upshot of the various criticisms of the neoclassical production function has been to call into question its very existence and all that follows from it. Consequently, it is more than a trivial theoretical point. In this article we will attempt to clarify the controversy somewhat by defining the

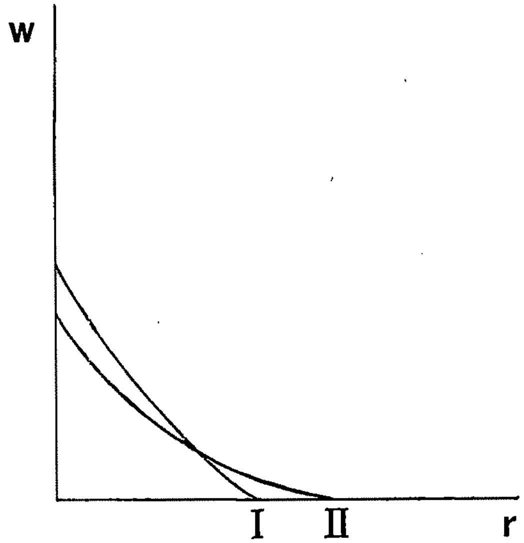


FIGURE 1

---

* Professor and associate professor of economics, respectively, Ohio University. We are indebted to Robert B. McGhee of the Electrical Engineering Department at Ohio State University and the late Charles Ferguson for key insights, to Michael Farren for his assistance in conducting computer simulations that suggested the direction of the inquiry, and to Hiroshi Ono for simplifying some of the mathematical argument. We also acknowledge the comments of Lionel McKenzie, Paul Samuelson, and Robert Solow as well as the encouragement of Joan Robinson whose cryptic injunction, "Do not bother. Neoclassical theory is no better off even when there is no reswitching," was a vital spur to further investigation.

[1] The best single summary of these attacks is provided by G. C. Harcourt. P. Sraffa was one of the first to raise this question. For a wide range of positions see the *Quarterly Journal of Economics* symposium consisting of Michael Bruno, Edwin Burmeister, and Eytan Sheshinski, P. Garegnani, David Levhari and Paul Samuelson, Michio Morishima, Luigi Pasinetti, and Samuelson as well as Charles E. Ferguson, Joan Robinson (1970, 1971a, b), and Robert Solow.

[2] For a summary of this argument, see Samuelson.

[3] This is the standard reswitching case found in Bruno, Burmeister, and Sheshinski, Garegnani, Morishima, Pasinetti, and Sraffa.

sufficient conditions for the existence of a well-ordered neoclassical production function. Simply stated, there is only one such condition, namely, that commodity prices be positive and finite for any positive value of the interest rate. We will proceed by first demonstrating that this condition will eliminate the possibility of reswitching and we will then argue that it also ensures that real Wicksell effects will dominate price Wicksell effects. Our analysis will be limited to the case of two-commodity indecomposable production systems.[4]

I. Characteristics of Factor Price Contours

Consider the following production system:

$$(1)\qquad A\begin{cases} a_0 = (a_{01} \quad a_{02}) \\ a = \begin{pmatrix} m_1 & m_2 \\ n_1 & n_2 \end{pmatrix} \end{cases}$$

where a_{01} and a_{02} denote the labor requirements for producing commodities one and two, respectively, and m and n represent the inputs of commodities one and two in the production of both themselves and each other.

Certain general constraints apply to the parameters of this production system. First, all parameters are assumed to have nonnegative values and, second, the parameters m_1 and n_2 must be less than unity since they are the quantities of commodities one and two required to produce themselves.

The price equations for this system are:

$$(2)\qquad 1 = (1 + r)(a_{01}w + m_1 + n_1p_2)$$

$$(3)\qquad p_2 = (1 + r)(a_{02}w + m_2 + n_2p_2)$$

where r denotes the rate of interest, w the wage rate, p_2 the price of commodity two, and commodity one is used as the measure of value.[5]

Expressions (2) and (3) can be manipulated to produce

$$(4)\qquad p_2 = \frac{a_{02} + (a_{01}m_2 - a_{02}m_1)(1 + r)}{a_{01} + (a_{02}n_1 - a_{01}n_2)(1 + r)}$$

At this point we introduce our assumption that commodity prices must be positive and finite for all positive values of r. In order for this assumption to hold, the expressions $(a_{01}m_2 - a_{02}m_1)$ and $(a_{02}n_1 - a_{01}n_2)$ must be positive. Otherwise, the numerator or denominator of (4) can be negative at some positive value for r and negative (or indeterminant) values for p_2 can result.[6] Thus, two basic additional constraints on the parameters of the production system are postulated:

$$(5)\qquad a_{01}m_2 - a_{02}m_1 > 0$$

$$(6)\qquad a_{02}n_1 - a_{01}n_2 > 0$$

The importance of these constraints will be seen shortly.

We turn now to the factor price contour itself. Equations (2) and (3) imply

$$(7)\qquad w = \{1 - (m_1 + n_2)(1 + r) + (m_1n_2 - n_1m_2)(1 + r)^2\} \div \{a_{01}(1 + r) + (a_{02}n_1 - a_{01}n_2)(1 + r)^2\}$$

which describes the factor price contour. Expression (6) requires that the denomina-

---

[4] This is for purposes of simplifying the discussion. Further, a good deal of the literature discussion is in terms of indecomposable systems. With slight modification a decomposable system with reswitching can be converted into an indecomposable one with reswitching.

[5] We implicitly assume that $p_1 > 0$. Some of the implications of $p_1 = 0$ are explored in fn. 10. Also, we have chosen to multiply all input costs by $(1+r)$ (rather than just capital costs), *à la* Garegnani. Multiplying just capital costs by $(1+r)$ will simplify things only slightly. For example, equation (4) below would be unchanged while equation (7) would merely have $(1+r)$ factored out of its denominator.

[6] For ease of discussion we will focus on price p_2. Actually, unless our constraints are satisfied, p_2 could appear to be negative simply because p_1 (the numeraire) had turned negative. We should also note that there is a possibility that both the numerator and denominator of (4) would turn negative at exactly the same point. Unless they were identical throughout this would produce an infinite commodity price in the vicinity of the point at which they became negative and an indeterminant price at that point.

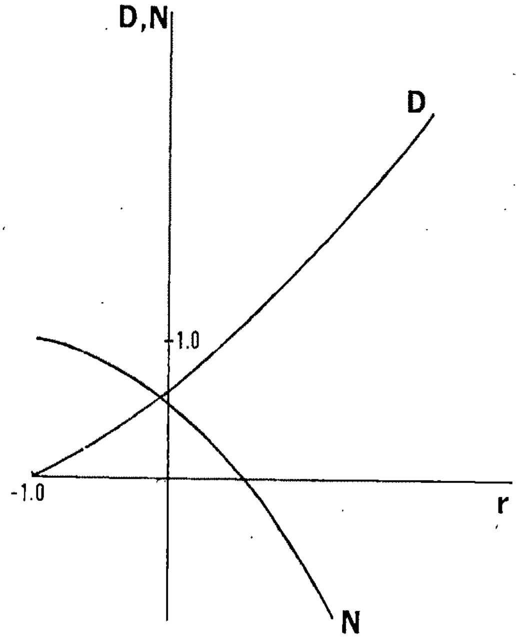


FIGURE 2

tor of (7) be positive at all times. Graphically, the denominator takes the form shown by curve *D* in Figure 2 and is monotonically increasing for positive values of $(1+r)$.

As to the numerator of (7), expressions (5) and (6) imply that[7]

$$m_1n_2 - n_1m_2 < 0 \tag{8}$$

and, in order for the factor price contour to have some positive values for both w and r, $(m_1+n_2)-(m_1n_2-m_2n_1)<1$. Consequently, the numerator of (7) is of the form indicated by the curve *N* of Figure 2. It is monotonically decreasing for positive values of $(1+r)$. Consequently, the factor price contour that results is monotonically decreasing for positive values of r.

At this point, we wish to emphasize that the forms of the numerator and denominator of (7) which are shown in Figure 2 are the only ones that will satisfy the assumption of positive and finite values for commodity prices for all positive values of r.

II. Remarks on Some Counterexamples

Michael Bruno, Edwin Burmeister, and Eytan Sheshinski, P. Garegnani, and Michio Morishima have postulated numerical counterexamples that purport to demonstrate the feasibility of reswitching among two-commodity indecomposable production systems.[8] However, an examination of their examples reveals that in all three cases at least one of the production systems they assume does not satisfy the assumption of positive and finite commodity prices for all positive r. In Table 1 parameters of these production systems and the respective values of the expressions $(a_{01}m_2-a_{02}m_1)$ and $(a_{02}n_1-a_{01}n_2)$ are presented.[9] Clearly, all of these examples violate the constraints we have imposed on the parameters of two-commodity indecomposable production systems. Thus, within our framework they do not demonstrate the possibility of reswitching.[10]

---

[7] Expression (5) indicates that $a_{01}>a_{02}m_1/m_2$ and (6) indicates that $a_{01}<a_{02}n_1/n_2$. Thus, $a_{02}m_1/m_2<a_{02}n_1/n_2$ and $m_1n_2<n_1m_2$.

[8] Levhari suggested a nonswitching theorem for such systems which prompted these examples. Pasinetti also presents a counterexample but it is generally agreed that his example is not a legitimate one in the case of two-commodity indecomposable systems.

[9] Table 1 presents the various counterexamples with a consistent notation. We have chosen to denote the common technique in each production system as technique 1. Some examples use it as technique 2. If the notation is reversed, the signs of the calculated values in the body of Table 1 will be reversed.

[10] In all fairness to the counterexamples it should be noted that p_1 and p_2 are positive within the range of values of r encompassing their switch points. However, at other values of r, either p_1 or p_2 is negative. Graphically, the full factor price contour implicit in these counterexamples *appears* to be one in which the wage rate proceeds to negative infinity as $(1+r)$ approaches the value which will produce $a_{01}+(a_{02}n_1-a_{01}n_2)(1+r)=0$ and then reappears from positive infinity beyond that value of $(1+r)$. However, wages are really not infinite. At the critical value of $(1+r)$, $p_1=0$. This can be shown by reversing the normalization (i.e., by using p_2 as the measure of value). Consequently, at this critical value of $(1+r)$ infinite wages are the result of dividing the actual wage by zero.

TABLE 1—PARAMETERS OF PRODUCTION SYSTEMS FOR VARIOUS RESWITCHING COUNTEREXAMPLES

	Author(s) of Counterexample		
Parameter	Garegnani	Bruno et al.	Morishima
Common Technique			
a_{01}	89/10	1.0	1.0
m_1	0	0.0	0.0
n_1	379/423	0.1	0.2
Other Techniques			
System I			
a_{02}	9/50	0.66	0.01
m_2	1/2	0.02	0.71
n_2	1/10	0.03	0.00
$a_{02}n_1-a_{01}n_2$	−0.7287	−0.2340	−0.4000
$a_{01}m_2-a_{02}m_1$	4.4500	0.0200	0.0100
System II			
a_{02}	3/2	0.01	0.2
m_2	1/4	0.71	1.0
n_2	5/12	0.00	0.0
$a_{02}n_1-a_{01}n_2$	−2.3643	0.0010	0.0100
$a_{01}m_2-a_{02}m_1$	2.2500	0.7100	1.0000

III. A General Nonreswitching Theorem

Failure of the various counterexamples of reswitching to meet the basic assumption of positive and finite commodity prices for all $r>0$ does not mean that reswitching does not exist. After all, there are an infinite number of counterexamples which could be constructed and perhaps there are some which both satisfy our constraints and show the presence of reswitching. However, we think not. What follows is a general proof that within the constraints we have established reswitching cannot occur. Assume two production systems that have in common the technique for producing commodity one. We note initially that there are two possible configurations of the denominators and numerators of the functions that define factor price contours (equation (7)) that are consistent with reswitching. These are shown in Figure 3. In panel A the numerator generated by one system (denoted as $f(r)$) is greater than that for the other ($f_1(r)$) for all values of $r>0$ which produce a positive value for either numerator. Now, if reswitching occurs, the factor price contour including $f(r)$ will dominate at both low and high interest rates. The denominator associated with $f(r)$ is $\phi(r)$ which is greater than $\phi_1(r)$ for all values of $r>0$.[11] This follows from two basic propositions. First, all denominators of factor price contours are equal to zero at $r=-1.0$

---

[11] There are actually eight different possible configurations of the numerators and denominators, four for the situation where $\phi(r)>\phi_1(r)$ (panel A) and four where $\phi(r)<\phi_1(r)$ (panel B). Case (i) is where $f(r)>f_1(r)$ for all r between 0 and r_0 (the r associated with $w=0$). In case (ii), $f(r)>f_1(r)$ for small values of r but $f(r)<f_1(r)$ for large values of r as r varies between 0 and r_0. Case (iii) is where $f(r)<f_1(r)$ for all r between 0 and r_0 and in case (iv), $f(r)<f_1(r)$ for small r and $f(r)>f_1(r)$ for large r as r goes from 0 to r_0. Panel A shows case (i) for $\phi(r)>\phi_1(r)$. With $\phi(r)>\phi_1(r)$, case (ii) would not be consistent with reswitching since w would not dominate at both w_0 (the wage associated with $r=0$) and at r_0; case (iii) is incapable of producing even a single switch point; and case (iv) is eliminated by assuming $w(0)>w_1(0)$. For those cases where $\phi_1(r)>\phi(r)$, case (i) will not yield any switchpoints; case (ii) will not produce dominance of w over w_1 at both w_0 and r_0; and case (iii) is inconsistent with $w(0)>w_1(0)$ and the existence of any switchpoints. This leaves only case (iv) which is shown in panel B. All this assumes that there are a maximum of two switchpoints between systems. In the type systems under discussion, this is the case.

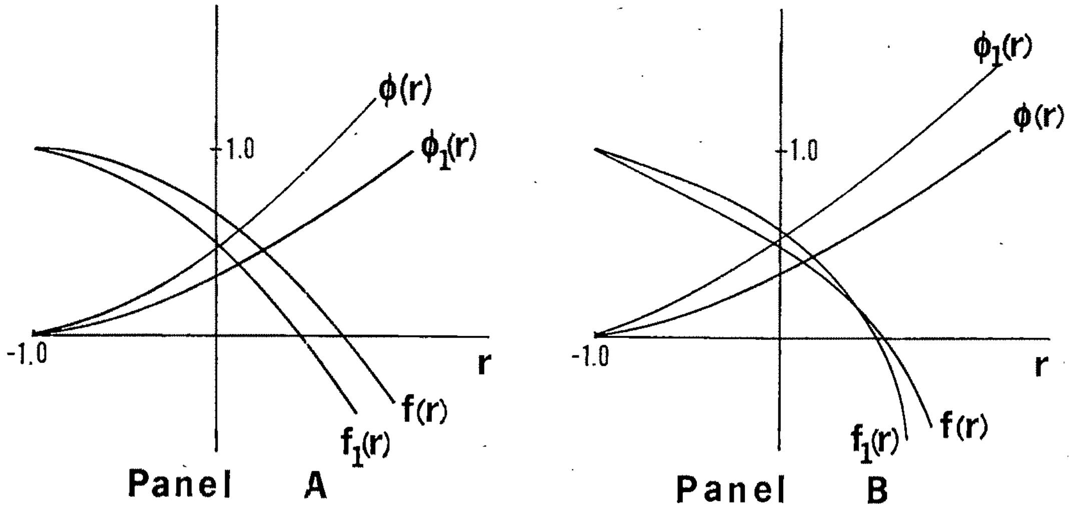


FIGURE 3

and for all $r > -1.0$ no denominators intersect one another. This can be seen by defining the denominators of the factor price contour functions in general terms as

$$\phi(r) = s(1 + r) + t(1 + r)^2 \tag{9}$$

$$\phi_1(r) = s_1(1 + r) + t_1(1 + r)^2 \tag{10}$$

where s and $s_1 = a_{01}$, $t = (a_{02}n_1 - a_{01}n_2)$, and the subscript 1 denotes a different production system.

Remembering that $s = s_1$, setting $\phi(r) = \phi_1(r)$ and solving shows that the condition for equality of $\phi(r)$ and $\phi_1(r)$ with $r > -1.0$ is $t = t_1$. But, this amounts to saying that $\phi(r)$ and $\phi_1(r)$ are identical. Consequently, no two denominators that differ from one another can intersect at $r > -1.0$. Second, the existence of a switch point implies

$$\frac{f(r_s)}{\phi(r_s)} = \frac{f_1(r_s)}{\phi_1(r_s)} \tag{11}$$

where r_s denotes the value of r at which a switch point exists.

Since $f(r) > f_1(r)$ at all points under consideration, $\phi(r_s) > \phi_1(r_s)$. But, if one value for $\phi(r)$ exceeds its counterpart in $\phi_1(r)$, $\phi(r) > \phi_1(r)$ at all $r > -1.0$. Therefore, we can state

$$s(1+r) + t(1+r)^2 > s(1+r) + t_1(1+r)^2, \quad r > -1.0 \tag{12}$$

which implies $t > t_1$.

Panel B illustrates the second general configuration that needs to be considered, viz., one in which the functions defining the numerators intersect one another in the positive quadrant. Again, $f(r)$ is associated with the factor price contour that must dominate at both low and high interest rates if reswitching occurs. In this case, however, $\phi(r) < \phi_1(r)$ for all $r > -1.0$. This follows from $f(r)$ at $r = 0$ being less than $f_1(r)$ at $r = 0$. Thus, we can write

$$s(1+r) + t(1+r)^2 < s(1+r) + t_1(1+r)^2, \quad r > -1.0 \tag{13}$$

which implies $t < t_1$.

We will now consider in detail the possibility of reswitching occurring in the case depicted in panel A of Figure 3. To do this we define the following

$$w = \frac{f(r)}{\phi(r)} = \frac{1 - j(1 + r) - k(1 + r)^2}{s(1 + r) + t(1 + r)^2} \tag{14}$$

and

$$(15)\quad w_1 = \frac{f_1(r)}{\phi_1(r)} = \frac{1 - j_1(1+r) - k_1(1+r)^2}{s(1+r) + t_1(1+r)^2}$$

where $j = m_1 + n_2$, $k = (n_1 m_2 - m_1 n_2)$, and the subscript 1 denotes a different production system.

Setting $w = w_1$ and simplifying yields:

$$(16)\quad (1+r)^2 + \frac{s(k_1 - k) + tj_1 - t_1 j}{tk_1 - t_1 k}(1+r) + \frac{s(j_1 - j) + (t_1 - t)}{tk_1 - t_1 k} = 0$$

For (16) to be consistent with reswitching it must have two positive roots. This requires that the coefficient associated with $(1+r)$ be negative and that the last term on the left-hand side of the equation be positive. Further, at this point we have established certain constraints on the parameters of (16). First, we know that $t > t_1$. Beyond this, if we evaluate $f(r)$ and $f_1(r)$ at $r=0$, $(1-j-k) > (1-j_1-k_1)$ which implies $(j+k) < (j_1+k_1)$. However, nothing is known at this juncture about the relationships between j and j_1 and k and k_1. Four possibilities need to be considered:

$$(17)\quad \begin{array}{cccc} j_1 > j & j_1 < j & j_1 < j & j_1 > j \\ k_1 > k & k_1 < k & k_1 > k & k_1 < k \end{array}$$

In the first case, $j_1 > j$ and $k_1 > k$, both the numerator and denominator of the coefficient associated with $(1+r)$ are positive, violating the necessary conditions for the existence of two positive roots for (16). Thus, this possibility can be disregarded.

The second possibility, $j_1 < j$ and $k_1 < k$, implies $(j_1 - j) < 0$ and $(k - k_1) > 0$. But we have established that $(j_1 + k_1) > (j + k)$ which requires that $(j_1 - j) > (k - k_1)$. Therefore, this case is inconsistent with our constraints and does not satisfy the conditions for the existence of reswitching.

The third case, $j_1 < j$ and $k_1 > k$, produces a positive denominator for the constant term of (16) and a negative numerator which again violates the conditions for the existence of two positive roots for equation (16).

The last case, $j_1 > j$ and $k_1 < k$, is the most complex of the panel A cases to analyze. We begin by noting that at any switch point

$$(18)\quad \frac{f(r)}{f_1(r)} = \frac{\phi(r)}{\phi_1(r)}$$

Turning first to the right-hand side of (18) we can define the nature of the relationship shown there by evaluating it at certain points. To begin, as r approaches -1.0 from $r=0$ the following holds:

$$(19)\quad \lim_{r \to -1.0} \frac{\phi(r)}{\phi_1(r)} = 1.0$$

Second,

$$(20)\quad \lim_{r \to \infty} \frac{\phi(r)}{\phi_1(r)} = \frac{t}{t_1}$$

Similar evaluations of $f(r)/f_1(r)$ yield the following:

$$(21)\quad \lim_{r \to -1.0} \frac{f(r)}{f_1(r)} = 1.0$$

$$(22)\quad \lim_{r \to \infty} \frac{f(r)}{f_1(r)} = \frac{k}{k_1}$$

At this point we define r_0 as being the value of r at which $f_1(r) = 0$. Now, as r approaches r_0 from below $f(r)/f_1(r)$ approaches ∞ and as r approaches r_0 from above $f(r)/f_1(r)$ approaches $-\infty$.

Graphic representations of $\phi(r)/\phi_1(r)$ and $f(r)/f_1(r)$ are shown in Figure 4 with $\phi(r)/\phi_1(r)$ being represented by the curve D and $f(r)/f_1(r)$ by the curve N. Curve D is a section of a rectangular hyperbola and increases at a decreasing rate and asymptotically approaches t/t_1. Curve N increases at an increasing rate from $r = -1.0$ through $r = r_0$ (at which point it ap-

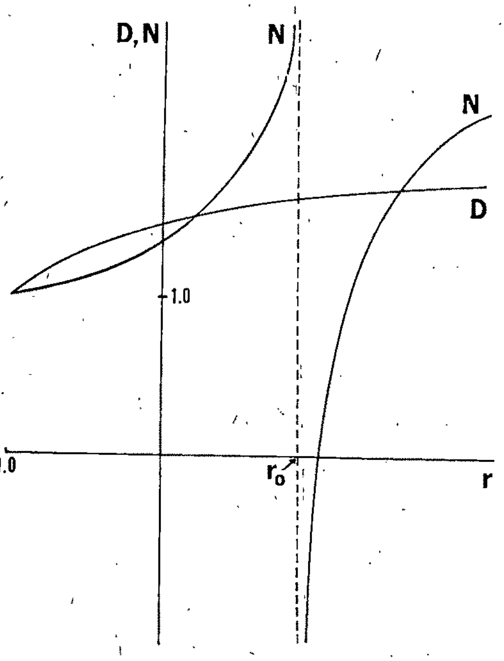


FIGURE 4

proaches infinity), reappears from negative infinity and increases at a decreasing rate and asymptotically approaches k/k_1.[12] Clearly, a necessary condition for a switchpoint is that D and N intersect. As drawn in Figure 4 they intersect at two points, indicating the presence of reswitching. This can occur as long as $k/k_1 > t/t_1$. However, the second switchpoint lies beyond r_0 and, therefore, is at a value of w (and w_1) which is negative. Thus, it lies outside the pertinent range of the factor price contour and does not represent an instance of reswitching that contradicts the basic neoclassical concept of switching. Another possibility is that $k/k_1 < t/t_1$. In this case no switchpoint occurs beyond r_0. However, it is only possible for one switchpoint to occur between $r = -1.0$ and $r = r_0$ due to the nature of the curves D and N. Thus, in this case the roots of (16) consist of either one positive and one negative, two negative, or two complex roots. Whichever is true, reswitching in the range $r = -1.0$ through $r = r_0$ is ruled out and the last possibility of reswitching occurring within the framework of the panel A assumptions is eliminated.

Turning now to the panel B configurations, we can establish the following constraints:

(23) $t < t_1,\ 0 < (1-j-k) < (1-j_1-k_1),$

and $(j+k) > (j_1+k_1)$

Again, there are the same four possible sets of relationships between j and k to be evaluated. In the first case, $j_1 > j$ and $k_1 > k$, we know that $(j_1 - j) > 0$ and $(k - k_1) < 0$. But, this violates the condition that $(j+k) > (j_1+k_1)$ which implies $(k-k_1) > (j_1 - j)$ and, therefore, this case cannot exist within our framework.

The second case, $j_1 < j$ and $k_1 < k$, produces negative values for both the denominator and numerator of the coefficient of $(1+r)$ in (16). Thus, it is positive, which violates the condition for the existence of two positive roots for equation (16).

The fourth possibility is that $j_1 > j$ and $k_1 < k$. In combination with $t_1 > t$ these conditions produce a positive numerator and negative denominator for the constant term in (16). Thus, there cannot be two

---

[12] The first derivative with respect to r of $f(r)/f_1(r)$ is a rational function with a numerator of quadratic form and a denominator that decreases at an increasing rate. The condition that $f(r)$ and $f_1(r)$ do not intersect in positive space is sufficient to insure that when the numerator is set equal to zero and solved for r the roots are complex. Thus, no minima or maxima for $f(r)/f_1(r)$ exist and the first derivative is positive for all values that satisfy the condition $-1.0 < r < r_0$. Further, since $f(r)$ initially is decreasing at a less rapid rate than $f_1(r)$, $f(r)/f_1(r)$ is increasing at an increasing rate in the vicinity of $r = -1.0$ despite the fact that in that vicinity the first-order term of the numerator (which is negative) dominates the second-order term (which is positive). Thus, in this vicinity the numerator of the derivative is actually decreasing but the denominator is decreasing more rapidly. As r increases, the second-order term of the numerator becomes relatively stronger and the numerator reaches a minimum and then begins to increase. In combination with the denominator's tendency to decrease at an increasing rate the result is a first derivative for $f(r)/f_1(r)$ that is monotonically increasing for the range $r = -1.0$ to $r = r_0$.

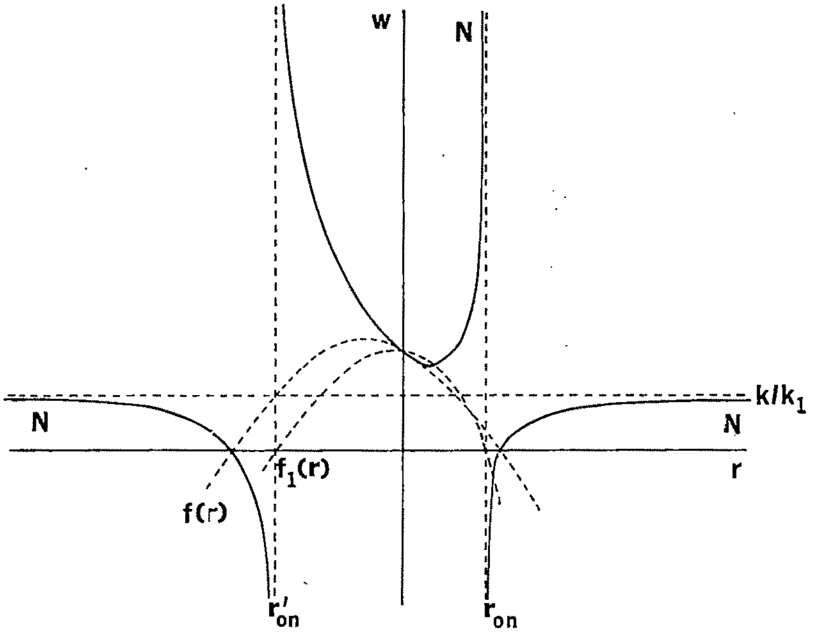


FIGURE 5

positive roots for (16) and reswitching is again ruled out.

This leaves only the third case, $j_1<j$ and $k_1>k$, which we have left till last because of its complexity. To deal with this case we will again resort to a consideration of the behavior of the expressions $f(r)/f_1(r)$ and $\phi(r)/\phi_1(r)$. Figure 5 displays graphically the behavior of $f(r)/f_1(r)$ over the full range of values of r for the panel B cases. The broken curves show the functions $f(r)$ and $f_1(r)$ while the solid curves (labeled N) depict the behavior of their quotient. The two critical values of r are those at r_{0n} and r'_{0n} which are those at which $f_1(r)=0$. As r approaches r_{0n} from below and as r approaches r'_{0n} from above, N approaches infinity in both cases. Moving from r'_{0n} to r_{0n}, N descends, passes through the point (1.0, -1.0), reaches a minimum at a positive value, and then increases again as r approaches r_{0n}. Beyond r_{0n}, N reappears from negative infinity, becomes positive, and asymptotically approaches k/k_1. At values of r less than r'_{0n}, N also reappears from negative infinity, becomes positive, and asymptotically approaches k/k_1 as r approaches $-\infty$.

Figure 6 shows the behavior of $\phi(r)/\phi_1(r)$ represented by the solid curves (D) with $\phi(r)$ and $\phi_1(r)$ being shown by the broken curves. Here, the critical values are those at which $\phi_1(r)$ is equal to zero (denoted by r_{0d} and r'_{0d}). At r_{0d} a discontinuity in the first derivative of D occurs. As r increases beyond r_{cd}, D decreases and asymptotically approaches t/t_1. On the other hand, as r decreases from r_{0d}, D again decreases, reaches a minimum, begins to rise, and approaches infinity as r approaches r'_{0d}. Beyond r'_{0d}, D appears from negative infinity, becomes positive, and asymptotically approaches t/t_1 as r approaches $-\infty$.

The relationships shown in Figures 5 and 6 are combined in Figure 7 to illustrate the fact that in all possible cases at least one root of (16) appears outside the

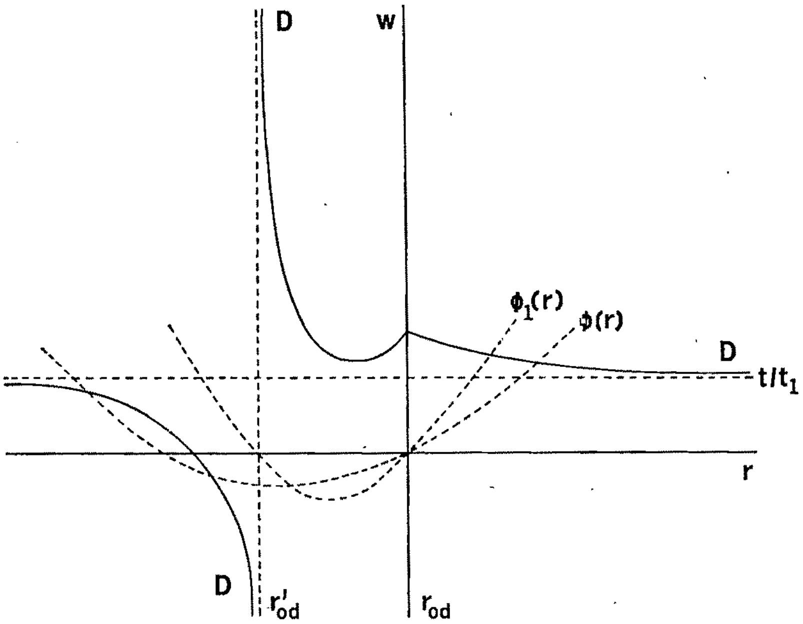


FIGURE 6

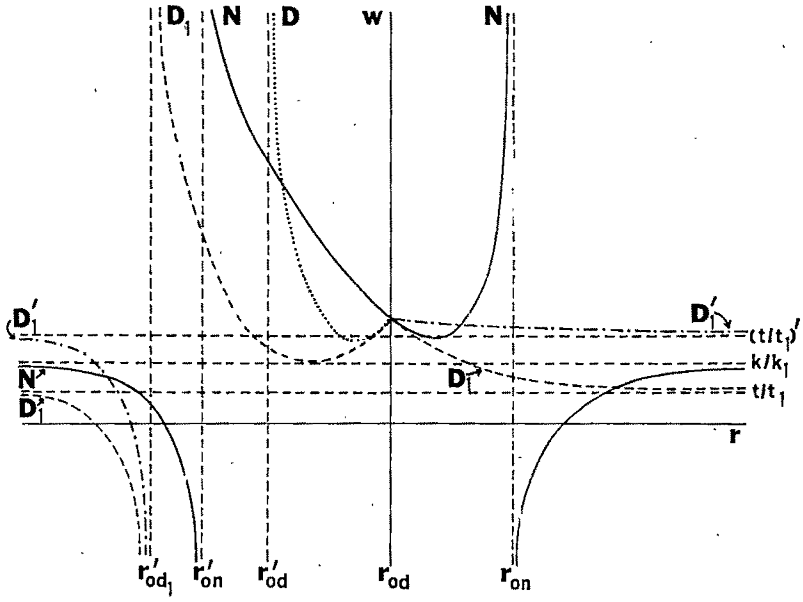


FIGURE 7

range $r=0$ to $r=r_{0n}$. This is sufficient to eliminate the possibility of economically meaningful reswitching which requires that both roots exist in the range $r=0$ to $r=r_{0n}$. We begin by observing the behavior of the curve D. It is drawn on the assumption that $r'_{0d}>r'_{0n}$. If this is the case, D must intersect N at some point between $r=-1.0$ and $r=r'_{0d}$. This is inconsistent with meaningful reswitching.

We next consider the possibility that $r'_{0n}>r'_{0d}$. This is shown by the curves labeled D_1 and D_1' in Figure 7. Clearly, the portion of D_1 lying between $r=-1.0$ and $r=r'_{0d}$ does not intersect N. At values of $r<r'_{0d_1}$, the curve D_1 is drawn to fit the case where $t/t_1<k/k_1$ and, again, no intersection occurs. However, if this is the case, the portion of D_1 that lies above r_{0d} must intersect N at a value of r that gives a negative wage rate, and reswitching is again ruled out. The other possibility is that at values of $r>r_{0d}$ the curve D_1' is the appropriate one and there is no intersection of D_1' and N in that range. However, if this is true, at values of $r<r'_{0d_1}$, D_1' will intersect N since $t/t_1>k/k_1$. Therefore, once again, the possibility of meaningful reswitching is eliminated.

IV. The Question of Wicksell Effects

We turn now to the question of the relationship between real and price Wicksell effects. As noted earlier, the critical question is the convexity or the concavity of the factor price contour with respect to the origin. For a well-ordered production function to exist, the factor price contour must be monotonically decreasing and decrease at a decreasing rate throughout the positive quadrant. Therefore, for all pertinent r, its first derivative must be negative and its second derivative positive. We have already defined the factor price contour in general terms as

$$w = \frac{f(r)}{\phi(r)} \tag{24}$$

where $\phi(r)>0$ for all positive r. Now,

$$\frac{dw}{dr} = \frac{\phi(r)f'(r) - f(r)\phi'(r)}{[\phi(r)]^2} \tag{25}$$

Now, as long as $f(r)$ is positive, i.e., the factor price contour lies in the positive quadrant, $\phi(r)f'(r)-f(r)\phi'(r)<0$. Since $[\phi(r)]^2>0$, this implies a monotonically decreasing factor price contour throughout the positive quadrant. To determine whether it decreases at a decreasing rate we will evaluate the second derivative over the range of values of $r>0$ that produce positive values for w. An examination of $f(r)$ reveals that as the value of $(1+r)$ increases from unity to $1/j$, the factor price contour passes completely through the positive quadrant and turns negative. Further, between $(1+r)=1$ and $(1+r)=1/j$, expression (25) is less than zero.

Now, let

$$\frac{dw}{dr} = \frac{f^*(r)}{\phi^*(r)} \tag{26}$$

Since $\phi^*(r)=[\phi(r)]^2$, it will be positive for all positive r while $f^*(r)$ will be negative for the portion of the factor price contour that lies in the positive quadrant. Differentiating,

$$\frac{d^2w}{dr^2} = \frac{\phi^*(r)f^{*\prime}(r) - f^*(r)\phi^{*\prime}(r)}{[\phi^*(r)]^2} \tag{27}$$

Obviously, $[\phi^*(r)]^2>0$. Expanding and simplifying $\phi^*(r)f^{*\prime}(r)-f^*(r)\phi^{*\prime}(r)$ reveals that it is positive if

$$\begin{aligned} 0 < kst(1+r)^3 - jt^2(1+r)^3 \\ + 3t^2(1+r)^2 + 3st(1+r) + s^2 \end{aligned} \tag{28}$$

All terms on the right-hand side of (28) are positive except for $-jt^2(1+r)^3$. However, if we substitute $1/j$ for $(1+r)$ in this term, it becomes

$$-t^2(1+r)^2 \tag{29}$$

When (29) is compared with the third term of (28), which is $3t^2(1+r)^2$, it is clear that expression (28) is positive for $(1+r)=1$ through $(1+r)=1/j$.[13] Thus, within the positive quadrant (remember that at $(1+r)=1/j$ the factor price contour has passed completely through the positive quadrant), the factor price contour is convex to the origin and the condition for dominance of real Wicksell effects is satisfied.

V. Conclusions

Our general conclusion is that within the framework of two-commodity indecomposable production systems the neoclassical production function is well-ordered as long as positive and finite commodity prices are assumed for all positive values of the interest rate.[14] Under these conditions, neither reswitching between techniques nor perverse Wicksell effects occur. What remains now is to explore the economic significance of the constraints we have imposed on the parameters of production systems in order to obtain these results.

REFERENCES

M. Bruno, E. Burmeister, and E. Sheshinski, "The Nature and Implications of the Reswitching of Techniques," *Quart. J. Econ.*, Nov. 1966, *80*, 526–53.

C. E. Ferguson, "Capital Theory Up to Date: A Comment on Mrs. Robinson's Article," *Can. J. Econ.*, May 1971, *4*, 250–54.

P. Garegnani, "Switching of Techniques," *Quart. J. Econ.*, Nov. 1966, *80*, 554–67.

F. H. Hahn and R. C. O. Mathews, "The Theory of Economic Growth: A Survey," *Econ. J.*, Dec. 1964, *74*, 779–902.

G. C. Harcourt, "Some Cambridge Controversies in the Theory of Capital," *J. Econ. Lit.*, June 1969, *7*, 369–405.

D. Levhari, "A Nonsubstitution Theorem and Switching of Techniques," *Quart. J. Econ.*, Feb. 1965, *79*, 98–105.

——— and P. A. Samuelson, "The Nonswitching Theorem is False," *Quart. J. Econ.*, Nov. 1966, *80*, 518–19.

M. Morishima, "Refutation of the Nonswitching Theorem," *Quart. J. Econ.*, Nov. 1966, *80*, 520–25.

L. L. Pasinetti, "Changes in the Rate of Profit and Switches of Techniques," *Quart. J. Econ.*, Nov. 1966, *80*, 503–17.

J. Robinson, "Capital Theory Up to Date," *Can. J. Econ.*, May 1970, *3*, 309–17.

———, (1971a) "Capital Theory Up to Date: A Reply," *Can. J. Econ.*, May 1971, *4*, 254–56.

———, (1971b) "The Measure of Capital: The End of the Controversy," *Econ. J.*, Sept. 1971, *81*, 597–602.

P. A. Samuelson, "A Summing Up," *Quart. J. Econ.*, Nov. 1966, *80*, 568–83.

R. M. Solow, *Capital Theory and the Rate of Return*, Amsterdam 1963.

P. Sraffa, *Production of Commodities by Means of Commodities*, Cambridge 1960.

---

[13] It may not be intuitively obvious to the reader why this is so. We are stating that if $3l^2(1+r)^2-jl^2(1+r)^3>0$ for $(1+r)=1/j$, it is positive for any $(1+r)\geqq 1$ but $<1/j$. Designate the value of r that makes $(1+r)=1/j$ as r_0. Now, simplifying the terms in question yields $3-j(1+r_0)>0$. By inspection, this clearly indicates that for values of $r<r_0$, the expression $3l^2(1+r)^2-jl^2(1+r)^3$ will be positive. Thus, the conclusion that the second derivative of the factor price contour is positive throughout the positive quadrant.

[14] What we have done can be interpreted in the framework of a Hahn-Mathews diagram. In the context of such a presentation, the parameters of a production system are shown as input-output vectors. The case we have been considering consists of two systems which have a common technique for producing one commodity. The input-output vectors for each system define its production possibility frontier. In the reswitching case the production possibility frontier of one system is dominated at both very low and very high values of r but dominates that of the other system over some range of intermediate values of r. Commodity prices in this type presentation are shown as normals to the production possibility frontier. Our constraints limit the possible configurations for these normals which in turn insures that the production possibility frontiers may not assume the positions necessary in order to produce reswitching.

The Social Costs of Input-Market Distortions

By DANIEL WISECARVER*

This paper examines the welfare effects of taxes (or other distortions) on factors of production. The analysis will demonstrate two major propositions: 1) Most importantly, given the usual *derived* demand curve for a productive factor, the social cost of a tax on that factor is correctly, completely, and most readily measured by the relevant area between the factor's demand and supply schedules. 2) Under variable proportions, the social cost that is so measured will entail two terms—corresponding precisely to the scale and substitution effects of derived demand theory—both of which represent true deadweight losses.

Although these propositions will be obvious to the practitioner of applied welfare economics, each has nevertheless been the source of considerable confusion in the recent literature. The first confusion is one of double counting,[1] as is explicitly illustrated in the recent paper by Ann Friedlaender (1971) in which the social cost of "value-of-service" (rather than marginal cost) pricing of *U.S.* rail services is estimated. Friedlaender's calculation consists of two distinct steps. First, the social costs that are measurable in the directly distorted rail markets are determined. Then, to this direct cost component is added a further element of deadweight loss that is alleged to arise from the fact that all goods requiring rail transport as an input are themselves nonoptimally priced—"indirectly taxed"—due solely to the distortion of rail prices.[2] We shall show, in Sections I, II, and III, that this sort of summing procedure is unequivocally wrong.

It might, however, be informative to speculate as to the source of this double-counting error. It is conceivable that the confusion has arisen from an eclectic survey of the two outwardly distinct approaches that have in the past been used to determine the social welfare costs of input-market distortions. The first approach, exemplified by the work of Albert Fishlow and Paul David, Jagdish Bhagwati and V. K. Ramaswami, and Harry Johnson, analyzes the problem within the context of the two-sector model of general equilibrium. It uses the concept of a social welfare function to depict the aggregate loss of welfare that is occasioned by an input-market distortion. This is defined to be the decrease in utility that society incurs when it is forced to move from an indifference curve which is tangent to the

---

* Assistant professor of economics, Ohio State University. The current version of this paper, completed while I was a staff member of the Office of Tax Analysis, U.S. Department of the Treasury, is a condensation of one chapter of my Ph.D. dissertation at the University of Chicago. I would like to thank and absolve A. C. Harberger, H. G. Lewis, and Larry Sjaastad for their extensive help and comments.

[1] This confusion also appears in Walter Oi and A. P. Hurter, Jr., pp. 340–41, although they do not attempt detailed calculations.

[2] Friedlaender's direct cost measure is calculated via $L_r = \frac{1}{2}\sum_i E_{ir} Z_{ir} \delta_{ir}^2$, where E_{ir} is the elasticity of demand for rail services for commodity type i, Z_{ir} is the proportion of total rail revenues received in i, and δ_{ir} is the distortion of the rail price of i. Similarly, her second cost component is $L = \frac{1}{2}\sum_j \epsilon_j \mu_j \gamma_j^2$, where ϵ_j is the elasticity of demand for commodity j, μ_j is the proportion of the economy's value-added that is generated by industry j, and γ_j is the price distortion for j caused by δ_{jr}, i.e., γ_j is the product of the cost share of rail services in j and δ_{jr}.

economy's undistorted transformation function to a lower social indifference curve which may or may not be tangent to the inferior transformation function that is dictated by the distortion. The second approach, exhibited by Albert Rees and Arnold C. Harberger (1964a, 1966), applies the current reincarnation of the theory of economic surplus directly to the distorted input markets themselves. This method treats the welfare cost of an input distortion as a simple money measure of the deadweight loss that the distortion imposes on the sum of the areas analogous to consumers' and producers' surpluses.

Among the apparent methodological differences between these approaches, one likely to receive considerable attention is the following: on the one hand, the surplus method treats the distortion as causing only one misallocative effect, that being the creation of a divergence between the input's demand and supply prices; on the other hand, the two-sector method explicitly analyzes two separate, distortion-induced effects. In Johnson's words:

> Of course, the total economic loss from factor market distortions can only be evaluated by taking into account, in addition to the inefficiency of production, the distortion of consumer choices due to the factor market distortion in producing a divergence of private from social opportunity costs in consumption. [p. 697]

Thus, even though these alternative approaches are applied to the same (qualitative) problem, our eclectic surveyor might question the compatibility of their results, feeling that the surplus approach captures only part of the total cost that is implied by the two-sector approach. In other words, should not the surplus method additionally allow for the fact that a (positive) distortion in the market for an input will cause the price of any product employing that input also to be set at an artificially high level, thereby imposing further deadweight losses? The correct answer is the negative. From the viewpoints of both derived demand theory and, at a significantly different level, the "distortion" approach to applied welfare economics, any attempt to add to the input-market measure of social cost a further element of loss that is alleged to occur in the output market when the supply price of the industry's product is forced to an (indirectly) distorted level will entail double counting of the actual costs involved.

That our second major point has been a source of confusion is exhibited by Richard Schmalensee's analysis of the welfare gains accruing from a fall in the price of a factor of production (of course, this is simply the obverse of our analysis of a tax which will raise the price of the factor in question) both within the market for that factor and within the market for the product that employs the factor. He finds, correctly, that under fixed proportions both measures of welfare gain are identical so that *either* measure could properly be used in a social accounting framework. However, when proportions are variable, the input-market measure exceeds that found in the product market by a term which is a pure substitution-in-production effect, as can easily be seen by suitable manipulation of Schmalensee's equation (7). From here, Schmalensee concludes that only the product-market measure should be used, contending that the factor-market measure will overstate the gain in welfare that is due to the decrease in factor price. Implicitly, therefore, he concludes that the "substitution" component of the factor-market measure does not represent true social gain. Sections II and III of the present paper will show that Schmalensee's conclusion is incorrect.[3]

---

[3] The recent, widely discussed study by Stephen P. Magee, p. 658, in which Schmalensee's faulty conclusion

In the analysis to follow, Section I refutes the double-counting error in terms of derived demand theory, utilizing the case of fixed proportions for ease of diagrammatical illustration. Section II, in treating the more complicated case of variable proportions, not only reaffirms our conclusion as to double counting but also shows that the substitution component of the factor-market measure of social cost represents a true element of deadweight loss. Finally, in Section III we review the entire analysis from the more intuitive standpoint of applied welfare economics.

I

The whole of the analysis to follow is based upon the standard, received theory of derived demand; factors of production do not, in and of themselves, yield utility to their demanders. Rather, inputs are demanded solely[4] because they can be used to provide those finished products that do yield utility to their purchasers. Thus, any utility that might be attributed to the area under a factor's demand curve is necessarily a mere reflection of the utility that its services ultimately provide. The word reflection here is very important, in that the utility garnered by the ultimate purchasers of the product of a factor's services *is* reflected in the demand curve for that factor. Obviously we do not claim that the total utility to be gained from the production of a good is the sum of the direct gain, as measured under that good's demand curve, and the reflected gain, as measured under the demand curve for factors used to produce the good. But then, by applying this obvious proposition in reverse, we have established a major point of this paper: it is incorrect to calculate utility *losses* in both the input *and* the output markets when only the input is explicitly taxed.

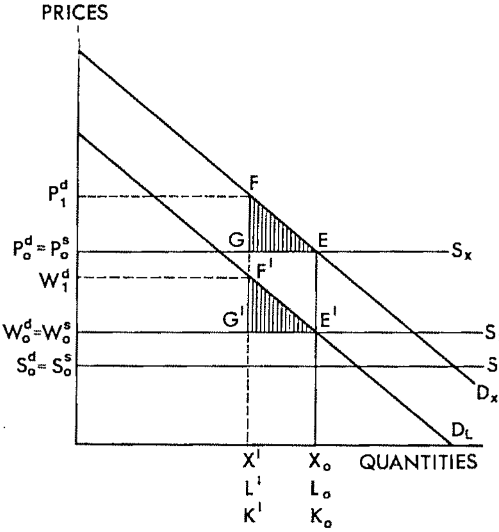


FIGURE 1

This conclusion is most easily seen when production technology is characterized by fixed proportions, an instructive, if not entirely realistic, framework of analysis. Hence, in the discussion to follow, we define units such that one unit of output (X) requires one unit each of labor (L) and capital (K), i.e., $X=\min(L, K)$. Consider Figure 1. The derived demand curve for L is constructed as the vertical difference between the demand for X and the supply of K, here assumed to be perfectly elastic as is the supply of L; the supply schedule for X is the vertical sum of the supplies of L and K. Overall equilibrium occurs where the demand and supply prices of X, L, and K equal each other, respectively, and where $P^s=w^s+s^s$ and $P^d=w^d+s^d$.[5] Thus

---

is favorably cited, gives further indication that this second source of confusion warrants clarification.

[4] It might be objected that certain inputs, particularly intermediate goods, are demanded in the aggregate partly for their productive services, but also partly for final consumption purposes. However, since the major point of this paper is that the best place for calculating the social costs of input taxes is the affected input markets themselves, this sort of objection can only strengthen our argument.

[5] In the notation of this section, P, w, and s are the prices of X, L, and K, respectively, and the superscripts d and s denote demand and supply prices. For a more

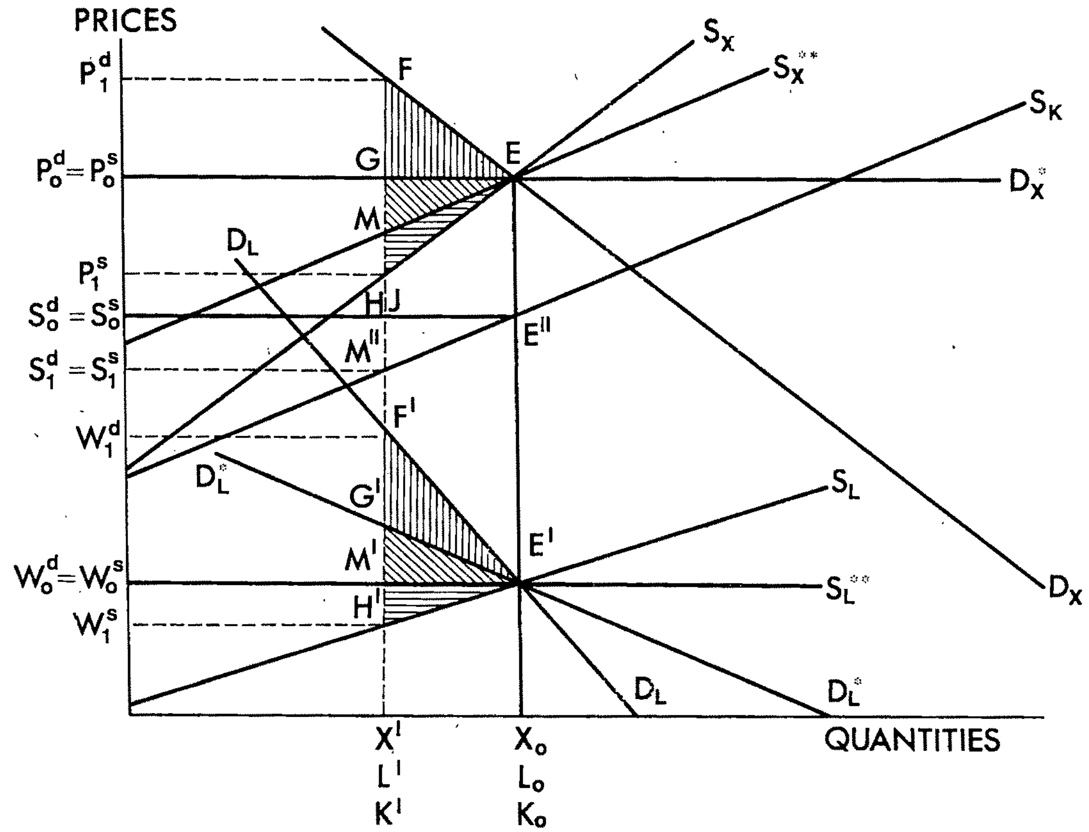


FIGURE 2

the points E and E' represent initial, undistorted equilibrium.

Suppose now that a tax of $F'G' = T$ per unit L is imposed, forcing equilibrium to shift to point F where $P_1^d = w_1^d + s_0^d = w_0^s + T + s_0^s$ and $P_1^s = w_0^s + s_0^s = P_0^s$. By the usual calculation of deadweight losses, the social cost of T in the market for L is $\frac{1}{2}T(L_0 - L') = E'F'G'$. If, however, one were to seek in the market for X a measure of social cost that might be attributed to the indirect distortion of P due to T, that measure would be $\frac{1}{2}(P_1^d - P_0^s)(X_0 - X') = -\frac{1}{2}T\Delta X = EFG = E'F'G'$. In the case depicted in Figure 1, where we assume that there are no rents to be lost, the deadweight cost of a tax can only be in terms of consumption utility—consumers' surplus—and this cost is equally well measured by EFG or $E'F'G'$. But to assert that the total deadweight cost is the *sum* of EFG and $E'F'G'$ is to engage in exact double counting.

This property of exact double counting also holds when rents to either or both factors are involved, as will be seen with reference to Figure 2. As before, D_L is constructed as the vertical difference between D_X and S_K, and S_X is the vertical sum of S_L and S_K (momentarily ignore the curves D_X^*, D_L^*, S_L^{**}, and S_X^{**}). Initial equilibrium at E is determined by the intersection of S_L and D_L at E'. If a tax of $F'H' = T$ is now imposed on labor, equilibrium shifts from E to F (E' to F') and

---

detailed description of the geometrical tool being used in this section, see Milton Friedman, ch. 7.

the resultant social cost is of course $-\frac{1}{2}T\Delta L = E'F'H'$. Again, due to the additive properties of demand and supply prices under fixed input proportions, $E'F'H'$ must equal the cost of T as it might be measured in the X market, i.e., must equal $-\frac{1}{2}(P_1^d - P_1^s)\Delta X = EFH$. For at F we have

$$P_1^d = w_1^d + s_1^d = w_1^s + T + s_1^s$$

and

$$P_1^s = w_1^s + s_1^s$$

so that

$$P_1^d - P_1^s = T$$

In order to show that the labor-market measure of social cost ($E'F'H'$) neglects no element that would require the inclusion of (any part of) the product-market measure (EFH), we shall break these triangles down to their component parts and see to what each component corresponds. To this end, it is convenient to consider the following hypothetical experiment. Suppose that the goal of the taxing authority is not the collection of revenue but is rather, for whatever reason, the restriction of the use of labor in X to L', thereby allowing T to vary inversely with the elasticities of S_L and D_L. Thus, given D_X, S_K, and S_L, the tax set is $F'H'$. Alternatively, if the demand for X were D_X^*, the derived demand for L would be D_L^*. If in addition the supply of L were S_L^{**} (implying a supply of X of S_X^{**}), then the tax would be set at $G'M'$ per unit labor. Now with these assumptions, there is obviously no consumers' surplus and no rent to labor to be lost, regardless of where we choose to measure social costs. However, there still is a deadweight cost of $E'G'M'$, which is easily shown to equal the areas EGM—producers' surplus as seen in the X market—and $E''JM''$, the rent, as seen in the K market, that is lost by owners of capital and yet is not transferred to any other economic entities. Thus, the area $E'G'M'$ in the directly taxed labor market captures the deadweight loss of rent that is imposed upon the untaxed factor.

If now the supply of labor is taken to be S_L, the authority will increase the tax on L to $G'H'$, which increase will entail no change in the position of capital owners. It does, however, increase the welfare cost by $E'M'H'$ at the expense of owners of labor. It is also clear that, with D_X^* and S_X, $P_0^d - P_1^s = G'H'$, so that $EGH = E'G'H'$ and therefore that $EMH = E'M'H'$.

Finally, let demand for X return to D_X and the tax to $F'H'$. This change does not further affect factor owners, but it does lead to a loss in consumers' surplus of EFG. And, since we have already shown that $EFH = E'F'H'$ and that $EGH = E'G'H'$, it follows that $EFG = E'F'G'$.

To summarize, the social cost of a tax on a factor of production in the case of fixed proportions (and when there are no other distortions in the economy) can be measured either in the directly distorted factor market or in the indirectly distorted output market; the market so chosen is a matter of indifference. Furthermore, we can divide the input-market measure into areas that are analogous to the traditional division of welfare cost into producers' and consumers' surpluses. The area above the taxed input's supply schedule and below its initial price—the analogue to producers' surplus—is rent lost by that input which is not gained by other sectors of the economy. The analogue of consumers' surplus, on the other hand, contains both the net social loss in rent to the untaxed factor[6] and the consumption utility lost because of the change in the price of the final good that is caused by the taxation of one of its inputs. The great error to be avoided is to measure tax-induced deadweight losses in the factor market and *then* add to these losses any element of further social cost that is alleged to stem from the product market.

---

[6] With fixed proportions, the argument can obviously be extended to n inputs. Then, the social loss of rent (corresponding to $E''JM''$ in Figure 2) to all of the $n-1$ untaxed factors will be counted in the consumers' surplus analogue below the taxed factor's demand curve (corresponding to $E'G'M'$ in Figure 2).

II

Although the case of variable proportions is more complicated, at least from the purely mathematical viewpoint, our basic conclusion with respect to the error of double counting still holds. Furthermore, the analysis is enriched because of the dual nature of the welfare effects of factor taxes; i.e., factor taxes entail both a loss of "reflected" utility due to the consequent product-price distortion and a waste of potential output due to the socially nonoptimal input mix that the tax will induce profit-maximizing firms to utilize.

Assume that in a competitive industry (composed of n identical firms) we have the linearly homogeneous production function[7]

$$X = F(L, K) \tag{1}$$

and that w and s, the prices of L and K, are fixed to the industry. Then it follows that[8]

$$\eta_L = \alpha_L \eta - \alpha_K \sigma \tag{2}$$

where η_L is the elasticity of derived demand for labor, the α terms represent the cost shares of the respective factors, η (a negative number) is the elasticity of demand for X, and σ (a positive number) is the elasticity of substitution between L and K. If, from a point of initial equilibrium, L is taxed at the percentage rate t, it follows that the resultant welfare-cost triangle is measured by

$$\begin{aligned} WC_L &= -\tfrac{1}{2} w_0 L_0 \eta_L t^2 \\ &= -\tfrac{1}{2} w_0 L_0 \alpha_L \eta t^2 + \tfrac{1}{2} w_0 L_0 \alpha_K \sigma t^2 \end{aligned} \tag{3}$$

Notice that these two deadweight terms correspond exactly to the scale (reflected utility) and substitution effects.

Now in order to determine the indirect social cost of t that *might* be measured in the market for X, we must know the effect of t on the supply schedule of X. By the condition of competitive equilibrium that price equal marginal *and* average cost, and by the properties of the assumed production function, it follows that

$$EP_x = \alpha_L Ew + \alpha_K Es \tag{4}$$

where E is the operator $d\ ln$. Fixed factor (supply) prices gives us $Es = 0$ and $Ew = t$, so that

$$EP_x = t_{xL} = \alpha_L t \tag{5}$$

where t_{xL} represents the indirect or implicit tax on X that is due to t. Then the implicit welfare cost of t that is brought about by the labor-tax-induced shift of the industry's supply schedule for X is

$$WC_x = -\tfrac{1}{2} P_{x0} X_0 t_{xL}^2 \eta = -\tfrac{1}{2} w_0 L_0 t^2 \alpha_L \eta \tag{6}$$

From (3) and (6), it follows that

$$WC_L = WC_x + \tfrac{1}{2} w_0 L_0 t^2 \alpha_K \sigma \tag{7}$$

Equation (7) shows us two important points. First, when factors can be substituted for each other (i.e., when $\sigma > 0$), WC_L exceeds WC_x by a term that is a pure "substitution-in-production" effect. It is this term that would be omitted if the social cost of t were calculated, as, for example, Schmalensee has suggested, solely in the output market. Indeed, in the extreme case of a zero elasticity of demand for X, a traditional "triangle-analysis" in the output market would lead to the incorrect conclusion that the social cost of a tax on labor employed in the X industry is zero. The fact of substitution implies that, even if there is no scale effect in the demand for labor, a tax on labor will induce an inefficient movement to a more capital-intensive production of X; the resultant inefficiency cost is captured automatically when the analysis is carried out within the taxed labor market itself.

---

[7] The same results obtain for any production function if the industry is in a position of long-run equilibrium. That is, in long-run equilibrium competitive firms operate at the minimum point of their average cost curves, and here the relevant production parameters are (locally) the same as those of constant returns to scale.

[8] See, for example, R. G. D. Allen and J. R. Hicks.

The second point emerging from equation (7) is that the labor-market measure of the tax's social cost also captures the "scale effect" when there is a scale effect, i.e., when $\eta \neq 0$. That is, WC_x is included in WC_L. Therefore, any attempt to add to WC_L the deadweight loss that one might conceive as occurring in the output market will entail an exact double counting of WC_x. So long as the relevant demand curve for labor is the usual derived demand, the welfare cost of a tax on labor is completely measured solely within the labor market.

There is no difficulty in interpreting the scale effect (WC_x) of WC_L: WC_x is simply the reflection of the deadweight loss of consumers' and producers' surpluses that is imposed on the economic agents in the X market when a tax on labor forces P_x to an artificially high level. However, the fact that the "substitution component" of WC_L—$\frac{1}{2}w_0L_0t^2\alpha_K\sigma$—is also a genuine element of social cost apparently requires further elaboration. There are at least three methods of explaining the mechanism that imposes this extra deadweight loss upon society. First, and easiest, we could simply appeal to the two-sector model of general equilibrium and assert that the substitution component of WC_L captures the essence of the inefficiency-in-production effect, while WC_x captures the distortion-of-consumer-choices effect (see the quotation from Johnson, above). However, even though this explanation has a great deal of intuitive appeal, it cannot withstand rigorous examination on its own merit, since our measure of WC_L was not explicitly derived from the general equilibrium model. Fortunately, the other two methods of explanation allow us to avoid any such "proof by assertion."

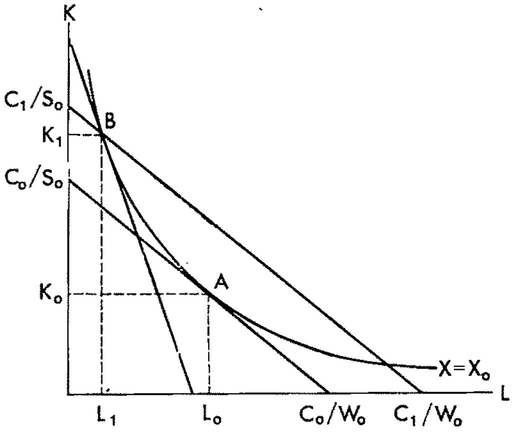


FIGURE 3

The second method relies on the fact that when technology allows input substitution, a tax placed on just one factor of production will lead profit-maximizing entrepreneurs to utilize an input mix which does not minimize social factor costs. For example, in Figure 3 we have the industry's isoquant map. Prior to the tax on labor, industry output X_0 was produced with the socially efficient input mix indicated by point A at total cost $C_0 = w_0L_0 + s_0K_0$. After the tax, X_0 would be produced with the input mix indicated by point B, implying a total *industry* cost of $w_0(1+t)L_1 + s_0K_1$. However, at point B private and social factor costs diverge. Since we are still assuming infinitely elastic factor-supply schedules, the factors' supply prices, and therefore their true social costs per unit, remain at w_0 and s_0; nevertheless, the tax has caused the industry to become more capital intensive than is socially optimal given w_0 and s_0. Therefore, the total social cost—total private

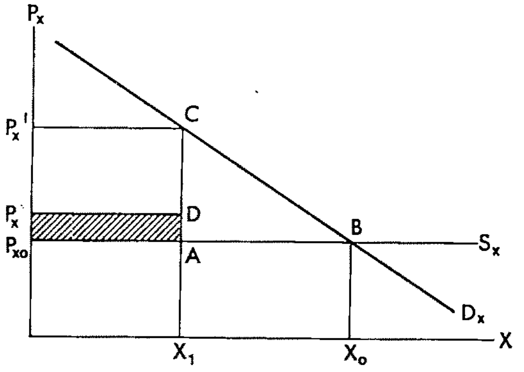


FIGURE 4

costs less tax payments—of X_0 when produced at input mix B exceeds C_0; i.e.,

$$C_1 - C_0 = w_0L_1 + s_0K_1 - w_0L_0 - s_0K_0$$

It can be shown that when $C_1 - C_0$ is approximated up to the quadratic terms of a Taylor series expansion about $t=0$, the resultant excess total factor cost is the substitution component of WC_L which is given in equation (7).

The key to understanding the third method—which is essentially a corollary of the preceding explanation—is to compare the total tax revenue that could be raised under two alternative taxation policies: (a) a tax on labor of t percent, and (b) a tax on X of $\alpha_L t = t_{xL}$ percent. That is, in this exercise we transform the tax t on labor into an implicit tax t_{xL} on X and compare the implied revenue (in terms of X) with the revenue (also in terms of X) that would be garnered if X were taxed directly at the percentage rate t_{xL}. The excess of revenue implied by policy (b) over that implied by policy (a) turns out to be $\frac{1}{2}w_0L_0t^2\alpha_K\sigma$, which is again the substitution component of WC_L.

The analysis is depicted in Figure 4. A direct tax on X of t_{xL} percent will raise the price facing consumers from P_{x0} to P_x' and will generate revenue of $P_{x0}P_x'CA$.[9] On the other hand, a tax of t percent levied on labor will cause the price of X again to rise from P_{x0} to P_x'; however, the factor tax alters the relative prices of labor and capital facing the industry. As can be seen by reference to Figure 3,[10] when the industry moves to more capital-intensive means of production the after-tax (but net-of-tax-payments) supply price of X exceeds P_{x0} at every level of output. Therefore, part of the area $P_{x0}P_x'CA$ which was tax revenue when X was directly taxed will become factor payments when labor is taxed instead, leaving the smaller area $P_x^*P_x'CD$ available as tax revenue. In other words, both tax policies induce the industry to increase the price at which it is willing to supply X from P_{x0} to P_x'. When X is taxed directly, all of the price increase, for each unit sold, becomes tax revenue. But when labor is taxed, part of the price increase reflects the higher supply price of X, at social factor cost, that results from the socially nonoptimal combination of inputs that the industry is induced to employ. Therefore, only part of $(P_x' - P_{x0})$ actually becomes tax revenue, and the area $P_{x0}ADP_x^*$ is the deadweight cost of producing X_1 at the socially nonoptimal input mix that results from taxing labor. Algebraically, this cost can be expressed as

$$\begin{aligned} P_{x0}ADP_x^* &= (P_x^* - P_{x0})X_1 \\ &= (P_x' - P_{x0})X_1 - w_0L_1t \end{aligned}$$

Expansion of the latter expression to the quadratic terms of a Taylor series yields our previously determined result.

---

[9] Notice that under fixed proportions the revenue $P_{x0}P_x'CA$ is exactly equal to the revenue that would be garnered from the tax t on labor, as measured in the labor market. These same qualitative results, i.e., that the welfare costs and revenue collections of t and t_{xL} (levied directly on X) are identical, obtains regardless of σ if $\epsilon_K = 0$. See the discussion of equation (16) below.

[10] The translation from Figure 3 to Figure 4 follows once it is recognized that for each level of output, i.e., for each isoquant, the tax on labor implies an excess social as well as private cost of production. And, since our analysis is one of comparative statics, price equals average cost both before and after the tax; P_x' is the tax-inclusive average cost, P_x^* the tax-exclusive average cost.

However, in spite of the fact that it is possible to represent all of WC_L in terms of the demand curve, supply curve, and (post-tax) cost conditions in the market for X, it should be evident, especially in light of the requisite mathematical manipulations, that practical measurement of the substitution component in the output market is intractable, at best. Furthermore, measures such as $ADP_x^*P_{x0}$ do not fit naturally into the framework of applied welfare economics since, among other reasons, this component of cost will exist whether the level of output changes or not (see Section III). Thus we see that the simple concept of the social cost of a distorted input market, as measured in the market for that input, not only captures all of the relevant deadweight components but also is the only concept that does so in a reasonably applicable manner.

The discussion so far has ignored any complications of factor rents. However, just as was the case under fixed proportions, none of our conclusions are altered when rents are included, i.e., when we assume noninfinite, factor-supply elasticities. Hence, only an outline of this more general case will be presented here. The entire analysis can be generated from the following four-equation model. Under constant returns to scale (or in long-run equilibrium), a competitive industry's demands for labor and capital may be written as

$$EL = \eta EP_x - \alpha_K\sigma E(w/s) \tag{8}$$

$$EK = \eta EP_x + \alpha_L\sigma E(w/s) \tag{9}$$

Also, we define ϵ_L and ϵ_K as the supply elasticities of L and K facing industry X, so that[11]

$$EL = \epsilon_L Ew \tag{10}$$

$$EK = \epsilon_K Es \tag{11}$$

By combining (9) and (11) to determine Es (having relied upon (4)) and substituting the result into (8), we can solve explicitly for the elasticity of derived demand for labor,

$$\eta_L = \frac{EL}{Ew} = \frac{\sigma\eta - \alpha_K\epsilon_K\sigma + \alpha_L\epsilon_K\eta}{\epsilon_K + \alpha_L\sigma - \alpha_K\eta}$$

Similarly, this simple factor-markets model, along with (4), gives us an explicit expression for the elasticity of product supply:[12]

$$\epsilon_x = \frac{\epsilon_L\epsilon_K + \sigma(\alpha_L\epsilon_L + \alpha_K\epsilon_K)}{\sigma + \alpha_L\epsilon_K + \alpha_K\epsilon_L}$$

Given η_L and ϵ_L, a rate of tax t on labor of course implies a welfare cost of

$$WC_L = -\tfrac{1}{2}w_0L_0t^2\left(\frac{\epsilon_L\eta_L}{\epsilon_L - \eta_L}\right) \tag{12}$$

Once again, to determine any implicit deadweight cost that occurs in the product market due to the imposition of t on L, we must calculate $EP_x = t_{xL}$ via (4). However, unlike the earlier case of equation (5), when t induces the industry to shift to more capital-intensive means of production, Es will not be zero, Ew will not simply be t, and t_{xL} will not in general be $\alpha_L t$. Consider Figure 5. Initial production of X_0 is accomplished with L_0 of labor and K_0 of capital when (w_0/s_0) represents relative factor prices (panel A). The tax on labor causes a new factor-price ratio, say $(w/s)'$. Now X_0 could still be produced at the factor combination A, implying a cost increase of w_0L_0t. Such a situation would emerge (in panel B) as a shift in the supply of X by $\alpha_L t$ percent (from S to S'). However, point A is no longer a cost-minimizing position for the industry; the tax will

---

[11] It should be noted that throughout we are assuming either that we are operating entirely on constant-elasticity curves or that the relevant parametric disturbances are small enough to make any changes in the relevant elasticities negligible.

[12] Both derivations here are almost identical to those found in Richard Muth.

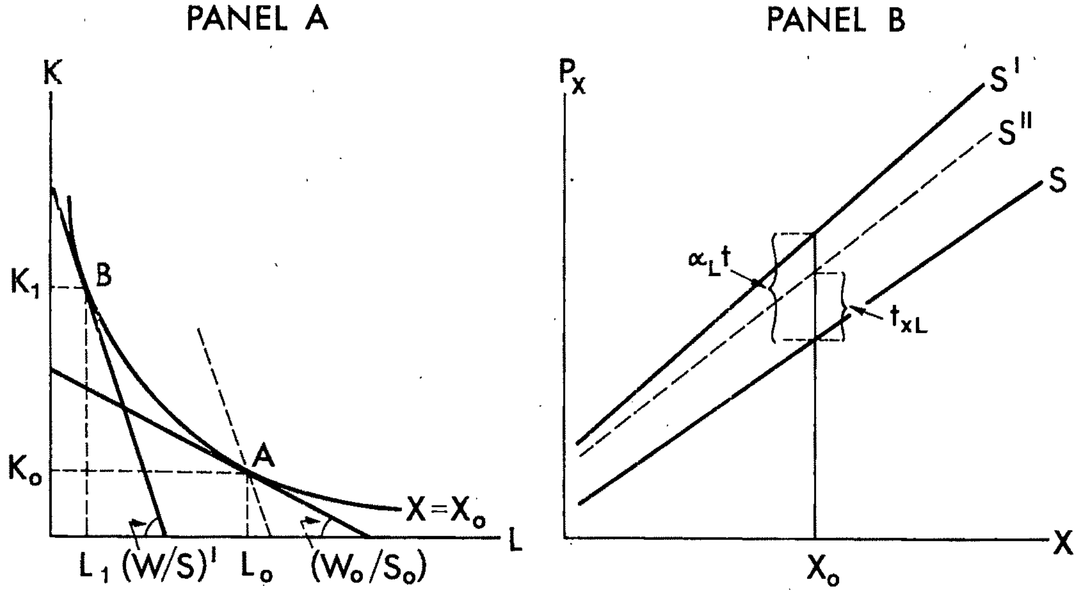


FIGURE 5

induce the new factor combination L_1 and K_1 at point B. Total industry costs will therefore not rise by $\alpha_L t$ percent but by some amount qualitatively represented by t_{xL} in panel B.

The simplest way to find the relevant percentage increase in industry supply price that is occasioned by t is to derive t_{xL} from output-constant factor-demand curves. This procedure has the virtue of yielding an expression that is independent of the elasticity of demand for X. Thus, in equations (8) and (9) we set $\eta EP_x = EX$ = zero. The resultant expressions, along with (11), give us the output-constant elasticity of demand for labor.

$$\overset{s}{\eta}_L = \frac{-\alpha_K \epsilon_K \sigma}{\epsilon_K + \alpha_L \sigma}$$

(here the superscript s denotes "substitution effect only"). In using (4) to define t_{xL}, we must recognize that P_x is the industry supply price, while the relevant w and s are the demand prices that the industry must pay for its inputs. When labor is taxed, the demand and supply prices of labor (w^d and w^s) differ by the absolute amount of the tax; therefore, to derive t_{xL} we must use the percentage change of w^d. Starting from a no-tax situation, it is easily shown that

$$(13) \qquad Ew^d \Big|_{X_0} = \left(\frac{\epsilon_L}{\epsilon_L - \eta_L} \right) t$$

Since there is no tax on capital, its demand and supply prices will always be equal; hence, $Es \,|\, _{X_0}$ is simply derived from (11) and the output-constant version of (9). This result and (13) give us

$$(14) \quad t_{xL} = EP_x$$

$$= \alpha_L t \cdot \left(\frac{\epsilon_L(\epsilon_K + \sigma)}{\epsilon_L \epsilon_K + \sigma(\alpha_L \epsilon_L + \alpha_K \epsilon_K)} \right)$$

Notice that $t_{xL} = \alpha_L t$, the case that always obtains with fixed factor prices, only when $\epsilon_L = \epsilon_K > 0$.[13]

---

[13] Notice also that $t_{xL} > \alpha_L t$ when $\epsilon_L > \epsilon_K$ and that $t_{xL} < \alpha_L t$ when $\epsilon_L < \epsilon_K$. While it might seem paradoxical to find that, when substitution is possible, the supply-price increase can actually exceed that which is possible when proportions are fixed, the puzzle is removed once we recognize that the individual firms within our com-

Thus, given t, ϵ_x and η, the implicit WC_x is

$$(15)\quad WC_x = -\tfrac{1}{2}P_{x0}X_0t_{xL}^2\left[\frac{\epsilon_x\eta}{\epsilon_x-\eta}\right]$$

Once (12) and (15) are expanded (and the denominator of WC_L is changed), a comparison reveals that

$$(16)\quad WC_L = WC_x + \tfrac{1}{2}w_0L_0t^2 \cdot\left(\frac{\alpha_K\sigma\epsilon_L\epsilon_K}{\epsilon_L\epsilon_K+\sigma(\alpha_L\epsilon_L+\alpha_K\epsilon_K)}\right)$$

Thus, (16) again shows that all of the deadweight loss of welfare that one might wish to measure in the market for X, due to its artificially high price, is captured in the directly taxed labor market when the labor-demand curve is defined in the usual manner, i.e., as a curve consisting of scale and substitution effects.[14] Furthermore, the substitution component of WC_L, which vanishes when $\sigma=0$, is independent of η and therefore of X. In addition, notice that when the untaxed factor (capital in the present exercise) is supplied to the industry with zero elasticity, the substitution component vanishes regardless of σ. This is simply a reflection of the fact that, for each level of output, there is only one combination of inputs that will be utilized when the amount of capital available is fixed.[15] Finally, all three of the arguments made above to verify that the substitution component of equation (7) is a genuine element of social cost that is measurable in the market for X also apply to the substitution component of (16).

III

Although the algebraic exercises of the previous sections clearly substantiate our two basic propositions, it is instructive to reconsider the entire problem from the more intuitive standpoint of applied welfare economics. To this end, we rely upon the following general formula for the welfare change that is associated with any policy variable z:

$$(17)\qquad WC = \int_{z=0}^{z^*}\sum_i D_i(z)\frac{\partial X_i}{\partial z}\,dz$$

where D_i is the absolute size of the distortion between demand and supply prices for activity i, and $(\partial X_i/\partial z)dz$ shows the change in the level of activity i as z is varied.[16] In our present application of this expression, z is the factor tax and all other D_i are assumed to be zero.

But this last assumption immediately proves our case with respect to the double-counting error. The only direct distortion is that in the factor market; while it is of course true that the factor tax causes a shift of the product's supply curve, this

---

petitive industry act as though they faced fixed factor prices. Therefore, when each of them demands more capital relative to labor, the outcome is to shift the industry's demand curve along a positively sloped supply of capital curve. The resultant increase in the price of capital will of course apply not only to the marginal units employed, but to all inframarginal units as well. For further elaboration of this line of analysis, see Friedman, ch. 5.

[14] With variable proportions, extension of the argument to the n-input case must as yet rely on the intuition of derived demand. Regardless of the number of inputs, it is simply incorrect to ascribe to a tax on an input utility losses in both the input and output markets. There is, in addition, some corroborating "circumstantial evidence." Following the model presented by W. E. Diewert, we can develop expressions for the elasticities of derived demand for factor n and for the supply of product X. As should be expected, ϵ_X contains terms in the elasticities of supply of all n inputs and in the partial elasticities of substitution between them. Also, η_n contains terms in the elasticities of supply of the $n-1$ other inputs, their partial elasticities of substitution, and the elasticity of demand for X. The circumstantial evidence is that these expressions (ϵ_X and η_n) reduce to equations that are presented in the text above for $n=2$. Unfortunately, even for $n=3$, the algebra required to carry out the same sort of analysis that would result in an expression analogous to (16) has proved to be prohibitive. (But see fn. 17.)

[15] This statement assumes that capital and labor are fully employed both before and after tax. Since any fiscal tool that led directly to unemployment would presumably be excluded from consideration the assumption appears to be reasonable.

[16] A more detailed exposition, of course, appears in Harberger (1971).

shift occurs in a market which is not itself distorted and hence entails no additional component of welfare cost.[17]

Furthermore, the whole of the measure summarized by equation (7) or (16) does in fact represent welfare change. All that expression (17) requires for its validity (at least in terms of our present discussion) is that the distortion represent the difference between freely determined demand price and true (social) marginal cost, i.e., supply price; the characteristics of the relevant activity that underlie and determine demand and supply prices are of no special significance. In particular, the fact that a factor's demand price reflects the possibility of substitution with other inputs in production as well as the demand price for final goods does not in any sense indicate, despite Schmalensee's contention, that some portion of this demand price—or some portion of the area under the factor's demand-price schedule—does not reflect true social value. Therefore, our second basic proposition that the substitution components of equations (7) and (16) are true elements of deadweight cost is shown to be valid as well.

In spite of all the above admonitions against attempting to account for the welfare effects of a factor tax in both the product and factor markets, there is one case in which such a procedure is legitimate. However, its illustration which follows also demonstrates perforce the pitfalls that await any such analysis.

Whenever equation (17) is applied to a distorted system, the effect of each ΔX_i (i.e., of each $(\partial X_i/\partial z)dz$) must be counted only one time. While this is usually assured by considering only *direct* distortions as being relevant, it can also be accomplished, if one insists on treating both t and t_{xL} (to continue the terminology of the previous section) as relevant distortions by removing the scale effect from the demand for labor. That is, this procedure requires that the deadweight cost of t be measured with reference to a substitution-effect-only demand for labor curve. Then, the total welfare cost of t would be correctly measured as, say, WC_x plus some WC_L^* which contained only a term in the elasticity of substitution, as, for example, the second term in equation (7) or (16).

However, we should notice that the practical application of even this "correct" separate accounting of both the explicit tax on labor and the implicit, derived tax on output would require a special degree of caution. If t ultimately results in a social cost of WC_x plus WC_L^*, the ΔX that we would observe in the product market is the proper term for the determination of WC_x. However, the ΔL that we would presumably observe in the labor market is not the proper term for WC_L^*. The observed ΔL will be the sum of the change in labor demanded at constant output, due to its tax-induced increase in price, and the change in labor employed that results from the shift of the constant-output demand curve for labor which will necessarily follow ΔX. Now since this latter demand shift occurs in a positively distorted market, it entails a "rectangle" of additional welfare cost, and this rectangle will be translated into a cost component that is exactly WC_x.

In order to see this last point, suppose that a given tax on labor is to be put into effect in miniscule steps such that in successive periods the existing tax is, for example, $.01t$, $.02t$, . . . , $.99t$, t. Then, after each piece of tax has been imposed, there will result demand and supply prices for labor that diverge by the then-existing, total amount of the tax. The locus of demand prices generated in this manner will reflect both the substitution away from labor, at each level of output, and the scale effect on output that will result from every successive addition to the demand price of labor. That is, this locus of labor's demand prices will in fact trace out the derived demand curve for labor. Thus, using the observed ΔL for the calculation of WC_L^* is equivalent to using labor's derived demand curve; and as we have seen, this procedure will, in conjunction with considering t_{xL} to be a relevant, separate distortion, overstate the social cost of t by precisely WC_x, despite initial efforts to avoid this pitfall of double counting.

---

[17] Notice that the reasoning in this paragraph is completely general, regardless of the number of inputs; therefore we need not worry about the algebraic difficulties encountered in fn. 14.

One further point should be made in closing. Although our analysis has been explicitly concerned only with the effects of just one distortion in one productive sector of the economy, our results will be fully compatible with models of general equilibrium and with any sort of generally distorted system. Because we have confined ourselves to the effects of a tax on the use of one input in only one activity, it is clear that the conclusion—i.e., that the relevant measure of welfare cost is completely captured by the deadweight-loss triangle below the derived demand curve in the input market itself—holds regardless of what that activity and/or input might be. It is therefore but a small step to conclude that our results must hold when we consider the welfare effects of a system of taxes that imposes differential rates upon different factor uses or upon different factors employed in the same activity, or of a system which contains both sorts of discriminatory elements.

Once we recognize the argument's generality, our analysis of the social costs of such a general system of distortions need proceed no farther than Harberger's model (1964a); for in that model Harberger explicitly recognizes that his "reduced form," reaction coefficients—$R_{ij}=\partial X_i/\partial T_j$, where the X_i represent the number of units of factor X used in activity i—" . . . reflect not only conditions of final demand and supply, but also conditions of factor substitution" (p. 68). We might also add that there is no reason for restricting ourselves to just two inputs which can be substituted for each other in production.[18] It has been the aim of this paper to clarify more explicitly the rationale behind the proper framework for analyzing the welfare effects of input-market distortions.

---

[18] By assuming that all inputs other than capital and labor are used in fixed proportions to their respective products, Harberger can express the social cost of an entire system of product and input taxes as $\Delta W=\frac{1}{2}\{\sum_i\sum_j G_{ij}B_iB_j+\sum_i\sum_j H_{ij}B_iE_j+\sum_j\sum_i M_{ji}E_jE_i+\sum_j\sum_i N_{ji}E_jB_i\}$, where B and E are taxes on the various uses of capital and labor, respectively, and $G_{ij}=\partial K_i/\partial B_j$, $H_{ij}=\partial K_i/\partial E_j$, $M_{ji}=\partial L_j/\partial E_i$ and $N_{ji}=\partial L_j/\partial B_i$. If we wish to allow for more than two inputs whose (partial) elasticities of substitution with each of several other inputs is nonzero, we can generalize Harberger's expression to $\Delta W'=\frac{1}{2}\sum_k^m\sum_l^m\cdot(\sum_i\sum_j F_{ij}^{kl}V_i^kV_j^l)$, where there are m variable inputs under consideration. Here, $F_{ij}^{kl}=\partial F_i^l/\partial V_j^k$ represents the change in the ith activity's use of input k due to a tax on input l in activity j. These changes are to be understood to allow for the rest of the sectors in the economy to adjust to their new equilibrium levels. For $m=2$ (K and L), $\Delta W'$ reduces to ΔW.

REFERENCES

R. G. D. Allen, *Mathematical Analysis for Economists*, New York 1938.

J. Bhagwati and V. K. Ramaswami, "Domestic Distortions, Tariffs, and the Theory of the Optimum Subsidy," *J. Polit. Econ.*, Feb. 1963, *71*, 44–50.

W. E. Diewert, "A Note on the Elasticity of Derived Demand in the N-Factor Case," *Economica*, May 1971, *38*, 192–98.

A. Fishlow and P. A. David, "Optimal Resource Allocation in an Imperfect Market Setting," *J. Polit. Econ.*, Dec. 1961, *69*, 529–44.

A. F. Friedlaender, *The Dilemma of Freight Transport Regulation*, Washington 1969.

———, "The Social Costs of Regulating the Railroads," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 226–34.

M. Friedman, *Price Theory: A Provisional Text*, Chicago 1962.

A. C. Harberger, (1964a) "The Measurement of Waste," *Amer. Econ. Rev. Proc.,* May 1964, *54,* 68–76.

———, (1964b) "Taxation, Resource Allocation and Welfare," in *The Role of Direct and Indirect Taxes in the Federal Revenue System,* Nat. Bur. Econ. Res. and Brookings Inst. conference report, Princeton 1964.

———, "Efficiency Effects of Taxes on Income from Capital," in M. Krzyaniak, ed., *Effects of the Corporation Income Tax: Papers Presented at the Symposium on Business Taxation,* Detroit 1966.

———, "Three Basic Postulates for Applied Welfare Economics," *J. Econ. Lit.,* Sept. 1971, *9,* 785–97.

J. R. Hicks, *The Theory of Wages,* 2d ed., New York 1963.

H. G. Johnson, "Factor Market Distortions and the Shape of the Transformation Curve," *Econometrica,* July 1966, *34,* 686–98.

S. P. Magee, "The Welfare Effects of Restrictions on U.S. Trade," *Brookings Papers,* Washington 1972, *3,* 645–701.

R. F. Muth, "The Derived Demand Curve for a Productive Factor and the Industry Supply Curve," *Oxford Econ. Pap.,* July 1964, *16,* 221–34.

W. Y. Oi and A. P. Hurter, Jr., *Economics of Private Truck Transportation,* Dubuque 1965.

A. Rees, "The Effects of Unions on Resource Allocation," *J. Polit. Econ.,* Oct. 1963, *71,* 69–78.

R. Schmalensee, "Consumer's Surplus and Producer's Goods," *Amer. Econ. Rev.,* Sept. 1971, *61,* 682–87.

Private Incentives and Public Information

By JOHN M. MARSHALL*

The traditional economics of information concerns itself with private incentives to seek knowledge of market opportunities. It typically avoids asking whether individual acts of gathering and using information lead to efficient resource allocation or whether such actions ought to be legally constrained. The question of efficiency also arises in the analysis of invention and in the investigation of speculative behavior. In spite of this problem's intrinsic interest and important implications, it has received relatively little attention.

Henry Manne, in his book on insider trading, views the ethical questions surrounding it as equivalent to the question of economic efficiency. He claims, in essence, that gains from trading on insider information provide an appropriate incentive for the managers of firms. This argument can be attacked either by questioning whether the redistribution of income involved in creating the incentives is ethical, as the traditional legal view would do, or by denying that Manne proves his contention, as Burton Malkiel does in his review of the book.

A second view is taken by Eugene Fama and Arthur Laffer, who distinguish between *trading information* and *operational information.* In concentrating on the production of trading information, they imply that it is a waste of resources, while operational information is not. This distinction is suggestive but not conclusive, for it leaves unresolved and undiscussed the question of gains from trading on operational information.

Jack Hirshleifer (1971) offers a resolution to this difficulty by arguing that the proper distinction is whether production can be affected by the information. If production cannot be affected, Hirshleifer employs the pure exchange economy as a model; in the alternative case the production and exchange economy is used. Hirshleifer advances the proposition that information in a pure exchange economy is socially worthless but privately valuable, leading to excessive incentives to produce it. In a production and exchange regime, information is socially valuable in this view, but again the incentives to produce it are often excessive.

While employing Hirshleifer's distinction, the present essay examines the problem from a different theoretical viewpoint. The problem of insuring the risks entailed in the receipt of information is shown to have an important bearing on the arguments and on their practical application. Familiar diagrammatic tools are used to illustrate the impacts of public and private information in general equilibrium, and the basic propositions are presented with stress on their generality and limitations. Implications of the theory which seem surprising at first are shown to be consistent with many aspects of observable behavior.

I. Pure Exchange

A. *News and Information*

In common usage the word information is ambiguous. It means either information that is known, as it is after being delivered, or unknown, as it is when it is purchased. The word "information" will be used here

---

* Assistant professor of economics, University of California, Santa Barbara. I am indebted for useful comments to Ben Bolch, Anthony Tang, Sandra Marshall, and the managing editor. Jack Hirshleifer and a reviewer made detailed suggestions, many of which were incorporated. I bear responsibility for errors that remain.

only in the latter sense, while the former meaning will be conveyed by the term "news." Thus, the purchase of *information* eventually results in *news*.

News is defined by the observation that when it arrives it alters individual beliefs. Although it is possible to imagine that news replaces inflexibly held beliefs with other inflexibly held beliefs, it is more natural to suppose that it alters probabilistic judgments about the future. The best context for viewing the impacts of such news is the state preference model of financial markets in which each conceivable outcome is represented by a state, s. The arrival of news modifies beliefs about the probabilities of states but leaves the list of possible states unchanged.[1]

In the state preference model a market for contingent claims is postulated. After trade in that market, each individual j knows the consumption Y_s^j he will enjoy in each possible outcome s. The set $\{Y_s^j\}_{s=1}^S$ is essentially the portfolio with which he faces the future.[2] His satisfaction with the portfolio is

$$\sum_{s=1}^{S} \pi_s^j U^j(Y_s^j)$$

where π_s^j is his subjectively held probability of event s, and $U^j(\ )$ is the utility of prospective consumption. All individuals are risk averse in the sense that the marginal utility of prospective consumption diminishes as prospective consumption increases.[3] In the pure exchange version of this model, the endowment of individual j in each state s is $\bar{Y}_s^j$, and his endowment portfolio (endowment point) is $\{\bar{Y}_s^j\}_{s=1}^S$. Assuming that trade is costless, equilibrium in the contingent contracts market produces an efficient distribution of portfolios in terms of individual *ex ante* preferences.[4] In this sense the inherent risks are efficiently insured at equilibrium.

In the special case of two individuals and two possible states, the equilibrium can be depicted by an Edgeworth box diagram (Figure 1). The point $\bar{Y}$ represents the endowment portfolios of both individuals. In this market they trade to point Y at a price ratio given by the slope of line m. The equilibrium point has the property that an indifference curve Ω_1 of individual one is tangent to an indifference curve Ω_2 of individual two at that point, and both indifference curves are tangent to the equilibrium price line. The contract curve is the locus of all tangencies of indifference curves, as usual.[5]

When news arrives the subjective probabilities π_s^j are replaced by different probabilities $\hat{\pi}_s^j$; this is, in essence, the impact of news. The arrival of news shifts the indifference surfaces of each individual because probabilities are parameters of the preference functions. If the market meets after the news arrives, it will move to an equilibrium Y' which is different from the equilibrium Y it would have found had the original beliefs about probability prevailed.

Figure 1 illustrates the case in which the news makes state one seem more likely than before. Individual one, whose endowment portfolio is relatively rich in state one goods, improves his consumptions for both states, while individual two loses in

---

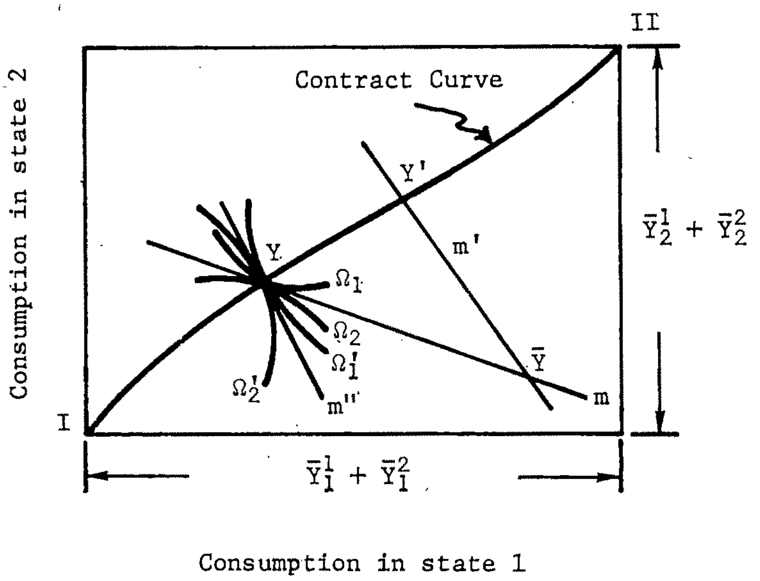
[1] Hirshleifer (1971) works in this context with the restriction that news always makes one state certain. Rather elaborate discussions of the plausibility and consistency of the assumptions are available. See Kenneth Arrow's original statement and the recent paper by Jacques Drèze. On the consistency of the assumed behavior given that subjective probabilities exist, see, for instance, Paul Samuelson. The existence of these probabilities, if not taken for granted, is treated by Leonard Savage.

[2] See Arrow and Drèze for the justification of ignoring the multiplicity of actual goods.

[3] See Arrow and Drèze on the importance and plausibility of this assumption.

[4] See Drèze for a detailed comparison of this concept with efficiency in the absence of risk.

[5] This diagrammatic technique is suggested by William Brainard and F. Trenery Dolbear, and by Richard Kihlstrom and Mark Pauly.



$\bar{Y}$	endowment point
Y	equilibrium achieved in the absence of news
Y'	equilibrium if news arrives before the market meets
Ω_1	indifference curve for individual one before news
Ω_2	indifference curve for individual two before news
Ω_1'	indifference curve for individual one after news
Ω_2'	indifference curve for individual two after news
m	price ratio in the absence of news
m'	price ratio if news arrives before trade
m''	price ratio if news arrives after trade

FIGURE 1. STABILITY OF THE CONTRACT CURVE

both states, relative to the situation that existed in the absence of news.[6] Thus, generally, the receipt of news affects equilibrium portfolios and the prospect of news implies a risk of altered consumption.

There is an important exception to the general rule which corresponds to the most tractable and best known special case. This is the case of pure exchange in which all individuals agree unanimously about probabilities both before and after the receipt of news, i.e., $\pi_s^j = \pi_s^{j'}$, and $\hat{\pi}_s^j = \hat{\pi}_s^{j'}$ for all j, j', and s. It is a special case because *the contract curve does not depend on probabilities, provided they are unanimously agreed upon.* The following paragraph will prove this assertion; its implications are clear. If the market is at equilibrium when the news arrives, there will be no change in portfolios when the market is reopened.[7] The indifference curves shift, but the equilibrium point of the previous market is already on the contract curve of the second market. In Figure 1, the new indifference curves Ω_1' and Ω_2' are tangent to each other at the old equilibrium Y. Trade from Y is impossible, although a new price ratio given by the slope of m'' is implied.

---

[6] If there is some variability of aggregate consumption over states, it is possible to define good news as that which increases mathematical expectation of aggregate consumption and bad news as that which decreases it. Then it is conceivable that a piece of good news would harm some persons through the redistributive effect or that bad news would benefit someone. An earlier version of this paper explored these possibilities in more detail.

[7] Contrast Hirshleifer (1971, p. 568).

Stability of the contract curve follows from examination of the formal efficiency requirements, which emerge as the first-order conditions to the following problem:

$$\text{(1)} \qquad \text{Maximize} \sum_{s=1}^{S} \pi_s^1 U^1(Y_s^1)$$

subject to:

$$\text{(1a)} \qquad U^{j*} = \sum_{s=1}^{S} \pi_s^j U^j(Y_s^j) \quad j = 2, 3, \ldots, n$$

and

$$\text{(1b)} \qquad \sum_{j=1}^{n} Y_s^j = \sum_{j=1}^{n} \overline{Y}_s^j \qquad s = 1, 2, \ldots, S$$

Here the second set of constraints says merely that one must allocate all the income in each state. The result of choosing parameters U^{j*} is to distinguish a single point on the contract curve.

At a solution to this problem there exist nonnegative constants μ_j, $j=2, 3, \ldots, n$, and λ_s, $s=1, 2, \ldots, S$, which fulfill conditions (2). Note that for symmetry μ_1 is defined as unity.

$$\text{(2)} \qquad \mu_j \pi_s^j U^{j'}(Y_s^j) - \lambda_s = 0 \qquad \text{all } j, s$$

The system is equivalent to

$$\text{(3)} \qquad \frac{\pi_s^1 U^{1'}(Y_s^1)}{\pi_1^1 U^{1'}(Y_1^1)} = \frac{\pi_s^j U^{j'}(Y_s^j)}{\pi_1^j U^{j'}(Y_1^j)} \qquad \begin{array}{r} \text{for } j = 2, 3, \ldots, n \\ s = 2, 3, \ldots, S \end{array}$$

This is a familiar efficiency condition: the marginal rate of substitution between any pair of contingent goods must be the same for all individuals. These conditions, together with the constraints on aggregate income (equation (1a)), determine the contract curve. Because probabilities are unanimously agreed, the ratio of probabilities is the same on each side of each equation in the system and may be canceled out. Thus, the contract curve is independent of the probabilities, and, consequently, news will not affect it in the case of pure exchange and probabilities agreed.

If probabilities are not unanimously agreed, however, the contract curve will depend upon them, and it will generally shift upon the receipt of news. Even if the market is in equilibrium when news arrives, there will be a readjustment of portfolios when it reopens. This makes the case of disagreement similar to the case of production and exchange, and essentially different from the case of unanimous agreement.

B. *The Value of Public Information*

The purchase of public information results in making the news known to all traders simultaneously. If information is purchased but the news is released only to one trader, it is private information. Private information in this sense is the inside information of Manne's treatment; Hirshleifer makes essentially the same distinction although he does not use a special term.

Writers who have discussed the social value of information explicitly (Hirshleifer) or implicitly (Manne, Fama-Laffer) are in agreement on one thing: when there are no production decisions which can be predicated on a piece of information, the information is socially valueless.[8] The logic of this position is nowhere completely clear. Hirshleifer demonstrates it only in a numerical example in which all individuals are identical in tastes, endowments, and subjective probabilities; others seem to take it for granted.

---

[8] An unpublished paper by Joseph Stiglitz makes valuable distinctions concerning the importance of production decisions and agreement about probabilities.

The conclusion is certainly true for an isolated individual; if he cannot affect the future, information concerning it is worthless except to satisfy idle curiosity. It is misleading to apply this argument to society, and it will be shown in Section IIc that the conclusion is false when individuals disagree about probabilities. In the case of unanimous agreement but differences in tastes and endowments, the conclusion holds, but the proof depends upon subtle arguments. It is worthwhile to present these arguments here not only to verify the essential principle, but also to show its implications for policy and to prepare the way for treatment of the case of production and exchange.

The relation of information to news is defined by considerations which are familiar from statistical decision theory. There is a probability space T, from which nature makes a random choice. Events s are subsets of T with probabilities π_s. Consider a partition of the probability space T into two mutually disjoint, collectively exhaustive sets A and B. Given A the conditional probability of event s is ${}_A\pi_s$, and given B it is ${}_B\pi_s$. The information that may be purchased consists of the knowledge whether the eventual outcome is in set A or set B.[9] For simplicity, one writes that the world is in "condition" A or B. No one knows which condition will actually apply, but it is assumed that all agree that the subjective probability of A is p_A and the probability of B is p_B, where the probabilities add up to unity. Consistency requires that $\pi_s = p_A\,{}_A\pi_s + p_B\,{}_B\pi_s$ for all s.

It is evident that the timing of the information is an important element in its value. If the information arrives after the market has closed, no reallocation of income can be made.[10] For a typical individual the expected utility of consumption in the absence of the information is $\sum_s \pi_s U(Y_s)$ and the expected utility in the presence of information is $p_A \sum_s {}_A\pi_s U(Y_s) + p_B \sum_s {}_B\pi_s U(Y_s)$ which is exactly the same thing. Because of this, information is valueless when it arrives after the market has closed.

Suppose that the information can be purchased before the market opens. Since the news has an impact on distribution, there is a risk involved in its receipt.[11] The social value of the information is highest when this risk is neutralized by holding another market prior to receiving the news. This market can take on a number of shapes, but the natural market is the one in which only the original S assets are traded. In this case the portfolios which individuals accumulate in the first market become their endowments when the second market is opened. It is convenient to speak of this prior market as *insuring the impacts of public information.* It permits traders to hedge against adverse news. The meaning of this insurance is clearer in the case of production, where the impacts of information can only be insured in a market which is distinct from the basic market. In the pure exchange case with probabilities agreed it can be shown (by arguments suggested in Section IIB) that efficient hedges are indeed created without involving any new assets. This preliminary market will be referred to as the hedging market.

Figure 2 indicates the trade that may, in principle, take place in the two markets. A typical individual starts with an endowment portfolio represented by the point $\overline{Y}$. In the hedging market he trades to the

---

[9] Irving Lavalle's treatment of the private value of information is definitive. This is a special case of what Lavalle calls "partition-J information." As far as I know, this simplification to a partition in two parts can be relaxed with no difficulty.

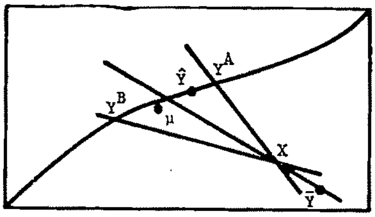
[10] Compare Hirshleifer (1971, p. 568). He regards this as the most realistic case.

[11] Hirshleifer (1971, p. 568) discusses in some detail the "distributive risk" which is referred to here.

TABLE 1

Situation	Information Not Expected	Information Expected	
First Market	**I-Market**	**Hedging-Market**	
Probabilities	π_s	$p_A,\ p_B,\ {}_A\pi_s,\ {}_B\pi_s$	
Endowments	$\overline{Y}_s$	$\overline{Y}_s$	
Equilibrium:			
Quantities	$\hat{Y}_s$	X_s	
Prices	$\hat{P}_s$	P_s^X	
Second Market	**No Second Market**	**A-Market**	**B-Market**
Conditional Posterior Probabilities		${}_A\pi_s$	${}_B\pi_s$
Posterior Endowments		X_s	X_s
Equilibrium:			
Conditional Posterior Quantities		Y_s^A	Y_s^B
Conditional Posterior Prices		P_s^A	P_s^B

point X, where he awaits the news. At the opening of the second market either condition A or condition B is known to exist. Trade brings the individual to a conditional posterior point Y^A or Y^B, depending upon the news announced. The points Y^A and Y^B are on the contract curve of the second market, which was already shown to be independent of the probabilities which obtain in that market. The market that meets if no information is purchased is called the I-market. In this market the point $\hat{Y}$ is the equilibrium point given the endowment $\overline{Y}$. This notation is summarized in Table 1.



$\overline{Y}$	initial endowment
X	proposed equilibrium
$\hat{Y}$	equilibrium when information is not expected
Y^A	conditional posterior portfolio for condition A
Y^B	conditional posterior portfolio for condition B
μ	alternative to portfolio X

FIGURE 2. TRADE BEFORE RECEIPT OF INFORMATION

The portfolios Y^A and Y^B are generally distinct points, but in the case of unanimous agreement about probabilities it can be shown that X is on the contract curve of the second market and, as a consequence, that Y^A and Y^B are identical to X and to each other. If X is not on the contract curve of the second market, each individual sees a portfolio of final consumptions $\{Y_s^A, Y_s^B\}_{s=1}^S$. Consider the distribution of portfolios represented by the point μ, which allows each individual in state s the conditional expectation, given s, of the payoffs in the portfolio $\{Y_s^A, Y_s^B\}$. That is,

$$(4)\quad \mu_s = \frac{p_A\ {}_A\pi_s}{\pi_s}\ Y_s^A + \frac{p_B\ {}_B\pi_s}{\pi_s}\ Y_s^B \quad \text{for all } s$$

The probabilities of the payoffs in the portfolio are $\pi_s = p_A\ {}_A\pi_s + p_B\ {}_B\pi_s$. These probabilities can be attained by an individual who holds portfolio μ provided he stands pat at μ in the A-market or B-market. This distribution of portfolios is preferred by every individual, because each individual is risk averse[12] and the portfolio is accessible to at least one individual (as illustrated in Figure 2).[13] In summary, if X is an equilibrium of the hedging market, it must be on the common contract curve of the A-market, B-market, and I-market.

---

[12] The probabilities associated with the portfolio payoffs Y_s^A and Y_s^B are $p_A\ {}_A\pi_s$ and $p_B\ {}_B\pi_s$. For each individual

When probabilities are unanimously agreed the equilibrium of the hedging market, X, is identical to the equilibrium of the I-market that would have resulted from the same initial endowments. Assuming that the prices which will prevail in the A-market and B-market are perfectly foreseen, the *ex ante* utility of the portfolio X is known and is

$$\begin{aligned}(5)\quad V(X \mid \{P_s^A\}, \{P_s^B\}, \{{}_A\pi_s\}, \{{}_B\pi_s\}, p_A, p_B) &= p_A U(X \mid \{{}_A\pi_s\}, \{P_s^A\}) \\ &\quad + p_B U(X \mid \{{}_B\pi_s\}, \{P_s^B\})\end{aligned}$$

where by definition

$$\begin{aligned}(6)\quad U(X \mid \{{}_A\pi_s\}, \{P_s^A\}) &= \max \sum_{s=1}^{S} {}_A\pi_s U(Y_s^A)\end{aligned}$$

subject to:

$$\sum_{s=1}^{S} P_s^A(Y_s^A - X_s) = 0$$

and similarly for $U(X \mid \{{}_B\pi_s\}, \{P_s^B\})$. This derived utility function V is convex in the usual sense.[14]

The equilibrium conditions for the X-market, namely

$$(7)\qquad \frac{P_s^X}{P_r^X} = \frac{V_s}{V_r} \qquad \text{all } r, s$$

where V_s is the partial derivative of V with respect to s, and

$$(8)\qquad \sum_{s=1}^{S} P_s^X(X_s - \overline{Y}_s) = 0$$

for every individual, are equivalent to the conditions

$$(9)\qquad \frac{P_s^X}{P_r^X} = \frac{\pi_s U'(X_s)}{\pi_r U'(X_r)} \qquad \text{for all } r \text{ and } s$$

together with the constraints. These conditions are identical to the conditions for

---

$$U(\mu_s) > \frac{p_A\, {}_A\pi_s}{\pi_s} U(Y_s^A) + \frac{p_B\, {}_B\pi_s}{\pi_s} U(Y_s^B)$$

and hence

$$\sum_{s=1}^{S} \pi_s U(\mu_s) > p_A \sum_{s=1}^{S} {}_A\pi_s U(Y_s^A) + p_B \sum_{s=1}^{S} {}_B\pi_s U(Y_s^B)$$

[13] The alternative μ must be accessible to some individual; if it is not, then

$$\text{(a)}\qquad \sum_{s=1}^{S} P_s^X \overline{Y}_s^j < \sum_{s=1}^{S} P_s^X \mu_s^j \qquad \text{all } j$$

Then summing over all individuals j yields

$$\text{(b)}\qquad \sum_{j=1}^{n} \sum_{s=1}^{S} P_s^X(\mu_s^j - \overline{Y}_s^j) \gneq 0$$

This is impossible because

$$\begin{aligned}\text{(c)}\quad \sum_{j=1}^{n} \mu_s^j &= \frac{p_A\, {}_A\pi_s}{\pi_s} \sum_{j=1}^{n} Y_s^{Aj} + \frac{p_B\, {}_B\pi_s}{\pi_s} \sum_{j=1}^{n} Y_s^{Bj} \\ &= \sum_{j=1}^{n} \overline{Y}_s^j \qquad \text{for all } s\end{aligned}$$

and as a consequence

$$\text{(d)}\qquad \sum_{s=1}^{S} P_s^X \sum_{j=1}^{n} (\mu_s^j - \overline{Y}_s^j) = 0$$

But the left-hand side of (d) is exactly the same as the left-hand side of (b) above, except for the order of summation, and equation (d) contradicts (b). Thus, the alternative μ is accessible to at least one party at the given price ratio.

[14] The indifference surfaces for the $U(X|A)$ and $U(X|B)$ are families of parallel planes with slopes dictated by $\{P_s^A\}$ and $\{P_s^B\}$, respectively. Let X^1 and X^2 be portfolios satisfying $U(X^1|A) \neq U(X^2|A)$, and let $X^0 = \theta X^1 + (1-\theta)X^2$, where $0<\theta<1$. Let Y^{A1} and Y^{A2} be the final consumption portfolios corresponding to X^1 and X^2. Since these are on the indifference planes of X^1 and X^2, $\theta Y^{A1} + (1-\theta)Y^{A2}$ is on the indifference plane of X^0. Then $U(X^0|A) = \sum_s {}_A\pi_s U(\theta Y^{A1} + (1-\theta)Y^{A2}) > \theta \sum_s {}_A\pi_s U(Y^{A1}) - (1-\theta) \sum_s {}_A\pi_s U(Y^{A2})$ by the basic assumption $U''<0$. The last expression in the string is equal to $\theta U(X^1|A) + (1-\theta)U(X^2|A)$. It can be similarly demonstrated that $U(X^0|B) > \theta U(X^1|B) + (1-\theta)U(X^2|B)$. It follows that the indifference surfaces of $V(X)$ are properly convex, for if $p_A U(X^1|A) + p_B U(X^1|B) = K$, and $p_A U(X^2|A) + p_B U(X^2|B) = K$ for some K, then $p_A U(X^0|A) + p_B U(X^0|B) > K$.

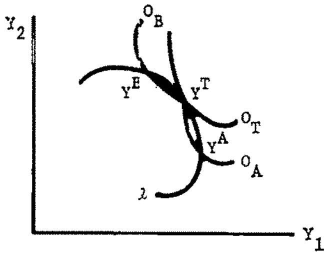
equilibrium in the I-market.[15] Geometrically this is interpreted as showing that the indifference surfaces of the derived utility functions V are tangent to indifference surfaces of the underlying utility functions U along the contract curve of the I-market. It follows that every hedging market equilibrium, X, is an I-market equilibrium $\hat{Y}$ which results from the same endowment, and vice versa.

The meaning of these results for the case of probabilities unanimously agreed can be stated briefly. If the impact of information is insured before its arrival, that insurance precludes further trade based on the news. After the meeting of the hedging market, the news changes nothing. The equilibrium of this market is exactly the one which would have occurred if trade had gone on in the absence of any prospect of information, and *ex ante* satisfaction of each individual is the same. In all, the information has no impact on distribution, no impact on *ex ante* satisfaction, and hence, no value.

In the contrary case when the news must arrive before its impact is insured in a preliminary market, the information is harmful. People can afford to pay something to suppress the news or to delay its arrival.[16] This is certain because given an opportunity to trade in advance of receiving the news, people do trade. No one stands pat off the contract curve. Public information in this situation is generally harmful; at best its impact can be counteracted by prior insurance.

C. *Private Information and Induced Public Information*

If information is purchased and the resulting news made known only to a single individual, it is private information. Private information is generally valuable, as may be illustrated without loss of generality by Figure 3. Opening trade at the equilibrium point Y^T, the individual faces an offer curve ℓ. His own offer curve is O_A in case of A, O_E in case of B, and O_T in case he does not purchase private information.



Y^A	equilibrium given A
Y^B	equilibrium given B
Y^T	equilibrium given T
ℓ	offer curve facing the trader
O_A	trader's offer curve in case of A
O_B	trader's offer curve in case of B
O_T	trader's offer curve in case of no information

FIGURE 3. THE VALUE OF PRIVATE INFORMATION

The three possible equilibria are Y^A, Y^B, and Y^T, in obvious notation. The trader's *ex ante* satisfaction in case of no information is $U(Y^T|T)$; in case he decides to purchase information it is $p_A U(Y^A|A) + p_B U(Y^B|B)$. An equivalent expression for $U(Y^T|T)$ is $p_A U(Y^T|A)+p_B U(Y^T|B)$. It follows that

---

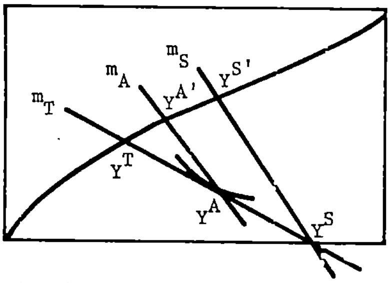
[15] The marginal value of a particular contract X is $V_s(X) = p_A U_s(X|A) + p_B U_s(X|B)$. Let the marginal utilities of income associated with the optima for the individual in the A-market and B-market be λ_A and λ_B. It is well known that $U_s(X|A) = \lambda_A P_s^A$, and $U_s(X|B) = \lambda_B P_s^B$. Among the necessary conditions for the maximum are ${}_A\pi_s U'(Y_s^A) - \lambda_A P_s^A = 0$, and ${}_B\pi_s U'(Y_s^B) - \lambda_B P_s^B = 0$. This means that $U_s(X|A) = {}_A\pi_s U'(Y_s^A)$ and $U_s(X|B) = {}_B\pi_s U'(Y_s^B)$. Substituting in the expression for $V_s(X)$ yields $V_s(X) = p_A\, {}_A\pi_s U'(Y_s^A) + p_B\, {}_B\pi_s U'(Y_s^B)$. In equilibrium each individual has $Y_s^A = Y_s^B = X_s$, with the result that $V_s(X) = \pi_s U'(X_s)$ for all s. Now if the price set is P_s^X, the equilibrium point X satisfies $P_s^X / P_r^X = V_s / V_r$ and $\sum_{s=1}^{S} P^X (X_s - \overline{Y}_s) = 0$ for every individual. This implies that $P_s^X / P_r^X = \pi_s U'(X_s) / \pi_r U'(X_r)$ for all s and r and for each individual, which, along with the constraints, are precisely the conditions needed for equilibrium in the I-market. It may be similarly shown that any equilibrium of the I-market is identical with an equilibrium of the hedging market. Because of the convexity of the derived indifference curves, any point which satisfies the necessary conditions for equilibrium in the hedging market is an equilibrium indeed.

[16] Hirshleifer also reports this conclusion (1971, p. 568).

$$(10)\quad \begin{aligned} p_A U(Y^A\mid A)+p_B U(Y^B\mid B)-U(Y\mid T) \\ =p_A[U(Y^A\mid A)-U(Y^T\mid A)] \\ +p_B[U(Y^B\mid B)-U(Y^T\mid B)] \end{aligned}$$

Both expressions on the right are nonnegative. Private information has positive value except in the case that both terms are zero, that is, in the case where the offer curve permits no adjustment to the information.

The individual who has purchased private information and made an adjustment to it then finds that publicizing the information will benefit him. If he is a small trader in a large market, the offer curve is the price line m_T in Figure 4. After he receives information that condition A applies he is at Y^A, assuming he does not speculate. If the information the individual has purchased can be publicly announced and verified, the individual can cause a shift in the price line from m_T to m_A. In this situation the entire market once again agrees upon the probabilities, and the optimum point for the individual is at $Y^{A'}$, which represents a further improvement for him. The motive to announce the information exists whether or not the owner of private information intends to speculate, and the result may be termed "induced public information."



Y^T	initial point
Y^A	optimum adjustment without announcement
Y^S	speculative position
$Y^{A'}$	optimum after announcement from Y^A
$Y^{S'}$	optimum after announcement from Y^S
m_T	initial price line
m_A	price line after announcement from Y^A
m_S	price line after announcement from Y^S

FIGURE 4. ANNOUNCEMENT AND SPECULATION

True speculative behavior is also possible if the information can be effectively announced. In that case, the possessor chooses a portfolio to hold at the time of announcement which will allow him the most advantageous situation at the prices his announcement will cause. This interim portfolio, Y^S in Figure 4, evidently lies to the southeast of Y^A along the offer curve, m_T. If the trader cannot sell short, Y^S is at the edge of the box as drawn, but if there are provisions to sell the contracts short in Y_2, a more extreme position is possible. More realistically, the offer curve bends inward at some point as drawn in Figure 3, effectively limiting the possibilities for speculative gain.[17] In Figure 4, $Y^{S'}$ represents consumption after readjustment; the gain from speculation is the difference in satisfaction between $Y^{S'}$ and $Y^{A'}$. The portfolio loss to the market at large is exactly the portfolio gain of the speculator.[18] Since public information is valueless, there is no benefit as a consequence of the announcement. Even neglecting the redistribution of income resulting from successful speculation, there is a social loss because resources are spent on producing the private information and an additional loss due to imperfect insurance of the impacts.

---

[17] In this case a simple tangency criterion illustrates the uniqueness of the optimum speculative portfolio. It seems unnecessary to spell that out here.

[18] Hirshleifer (1971) recognizes the incentive to announce information in the pure exchange economy and correctly attributes to it a part of the excessive incentive to spend on information in that case. Manne and Fama-Laffer do not treat announcements directly. The rebalancing of portfolios after the announcement of information, i.e., the move from Y^S to $Y^{S'}$, is an aspect of speculation which seems to have been overlooked by Hirshleifer (1971, 1972), by Manne, and by Fama-Laffer. It arises naturally from the theory; whether it has practical importance is an interesting question.

It should not be taken for granted that every possessor of private information will be able to speculate. The information may be incapable of being announced or it may be impossible to prove. On the other hand, many kinds of information can be announced, and advance information about the contents of a governmental or corporate policy statement is automatically made known.

Secrecy has an important bearing on the value of information; the more secret it is, the more valuable. Even when the news satisfies all the criteria for successful announcement, it has no impact if the market has already discounted it. Most traders will not know exactly what considerations and what information have gone into making market prices. Lacking this insight, the purchaser of private information can make the initial gain from adjusting his own portfolio, but he cannot be sure of making the additional gains from announcement. The value of information depends critically upon the degree to which it is secret.[19]

In practice the market is an important transmitter of information. This implies that the private value of the information depends upon the extent to which it is automatically spread to the market by the possessor's trading behavior. If the possessor is a small trader in a large anonymous market, he can exploit the information very effectively; but if he is a bellwether, other traders may follow his moves so closely that his speculative gains vanish. This factor in the private value of information can be termed "disguise." The greater the ability to disguise trading behavior, the greater the value of private information.

Quite apart from the problem of disguise, a small market is a disadvantage for the possessor of private information. Consider just two traders, one of whom has the information, while the other is ignorant of it. The owner of the information faces an offer curve that is rather sharply bent, reflecting the risk aversion of his trading partner. His adjustments are greatly restricted by this fact; both his initial shift and the possible gains from announcement are limited. In such a case the unwarranted expenditure of resources upon socially worthless information is a lesser problem than in the case of a large anonymous market.

D. *Implications*

The argument has been that public information is valueless and private information valuable, leading to an inefficient allocation of resources by overspending on information. This might imply a policy of suppressing these kinds of information, but does it? The logic is compelling, given the definitions, but it seems a paradox that more information should be bad instead of good, and a policy of suppressing information is antithetical to western ideals of political economy. There are two lines of argument leading out of this quandary, each of which exposes a different aspect of the problem.

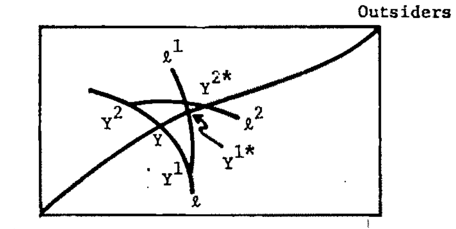
One approach starts with two groups of traders: insiders who have the information and outsiders who do not. Each group is homogeneous and trades as a unit. Insiders are depicted as occupying the lower left-hand corner in Figure 5. When the market opens insiders begin a speculative gambit, trading either to Y^1 or Y^2, announcing the information, and rebalancing back to Y^{1*} or Y^{2*}. Outsiders do not know which direction the speculative maneuver will follow, nor are they aware, even after it has begun, which direction it is following. They only know that either way they suffer a decrease in each portfolio asset. Outsiders have three alternatives for reducing their losses; they may

---

[19] The question of secrecy involves rather complex information in itself. Conceivably a person would purchase the primary information from one source and an assessment of its secrecy from another. This creates a complex set of decisions that deserves study in its own right. All that can be said here is that these expenditures on producing the secondary information also represent inefficient use of resources within the present model.

1) enter into trade without the information,
2) purchase the information, then enter into trade, or
3) refuse to trade with the insiders.

The third alternative, exclusion, is the best policy for the outsiders provided it is possible, for it keeps the distribution of assets fixed at point *Y* of Figure 5. Outsiders who stand pat until prices settle in the wake of the speculative gambit completely avoid loss. Unfortunately for outsiders, exclusion is seldom a practical strategy. In a more realistic model where trade and information gathering are continuous and where individuals use financial markets to save and dissave, exclusion could not be employed by all outsiders. Some outsiders would always be in the process of purchasing or liquidating portfolios, and they would have no way of knowing when they were vulnerable to announcements and subsequent shifts. Outsiders are not only ignorant of the content of the news insiders get, outsiders do not even know when the speculative gambit is over.



Y	initial equilibrium
Y^1Y^{1*}	speculative gambit if news favors state 1
Y^2Y^{2*}	speculative gambit if news favors state 2
ℓ	outsiders' offer curve before announcement
ℓ^1	outsiders' offer curve after announcement if news favors state 1
ℓ^2	outsiders' offer curve after announcement if news favors state 2

FIGURE 5. INSIDERS VERSUS OUTSIDERS

Exclusion being impossible, the choice remains between alternatives 1) and 2). Alternative 1) is to accept the losses imposed by the insiders, while alternative 2) means blocking the speculative ploy by informing all outsiders of the news. In alternative 2) the outsiders do not remain at point *Y* for they suffer the expense of purchasing the information, which is not illustrated in Figure 5. In principle their expenditure causes the entire box to shrink, just as the expenditure by insiders reduced its size.

Assuming that the cost of information to outsiders is the same as it was to insiders, and that insiders were wise in making the expenditure on the information, the choice of strategy is simple. Diminution of insiders' portfolios due to expenditures on the information was less than the expected gains from speculation. In this case outsiders are also wise to purchase the information collectively, for their expenditure on the information is less than the expected portfolio loss they would experience from the insiders' speculation. On the other hand, if insiders have overspent on information, it will not pay outsiders to spend on it. It is cheaper to take the losses which insiders are going to inflict.[20]

Whether the cost of information is greater or less to outsiders is an open question. The knowledge that insiders have the information may contain clues which make it cheaper for outsiders to obtain it. Outsiders could afford to pay any single insider for revealing the information instead of speculating on it. The traitorous insider might then reap more than his share of the speculative gains. If he bargained successfully, he would gain as much as all the insiders together could have made in speculating. The difficulty of verifying the information probably makes this kind of transaction generally much less remunerative than the argument suggests, but the strategy of the insider who sells information to outsiders is still competitive with the speculative strategy, especially for an insider with small assets. On the other hand, the cost of disseminating the information may be burdensome to outsiders. Finally, what of the costs involved in keeping track of the insiders? The fact that insiders have information is in itself an important piece of news. Outsiders must be wary of bluffs and bubbles. All of these factors make alternative 1) a more likely solution to the problem.

---

[20] The argument is strictly correct only when the gains are relatively small. It should be noted that outsiders do not know which way speculation will go, so their loss is probabilistic.

Alternative 1) is also the right solution when the possessors of information are well known. In this case there is no need for outsiders to buy the information; they simply keep close watch on insiders' trading and imitate it. This strategy reduces the range for speculation and prevents insiders from making important gains. Undoubtedly there are costs associated with monitoring the activity of insiders, just as there are costs in finding out the information directly. In practice, groups of insiders continuously discover new pieces of information and outsiders make a continuous series of decisions to buy it, ignore it, or monitor insiders' behavior.

Seen in this light, the theory is consistent with many aspects of reality. Rather than suppressing information, outsiders buy information and watch insiders' moves. Outsiders are acting rationally in doing this, and, given that insiders have private information, this expenditure can be efficient for outsiders.

Much the same result is reached by another route. The high cost of preventing individual inquiry or eliminating announcements is a deterrent to any strategy of suppressing information. Direct efforts to suppress information may backfire because they heighten its private value. Since the goal is to reduce the net incentives for private production of unproductive information, there is an alternative to suppression. The alternative focuses on reducing two of the conditions that make private information valuable, namely the secrecy of privately held information and the disguise with which insiders operate. The way to reduce secrecy is not to suppress information but to seek it out and publish it; the way to reduce disguise is to monitor markets and the main traders in them. This involves some expense, but it reduces incentives for insiders to spend on information and may represent a net improvement. Freedom of inquiry and freedom of the press may have a role in reducing losses from excessive spending on private information. Obviously, this is not an attempt to define the optimal strategy for reducing inefficient expenditure on information. It simply presents the rationale behind a policy of permitting and perhaps subsidizing the production of public information.

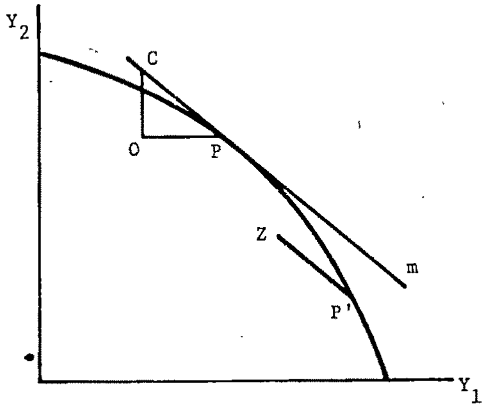
II. Production and Exchange

A. *News*

This section extends the analysis of information to a regime involving production. Following Debreu, it assumes that productive activity cannot affect the probabilities of outcomes but that it is possible to transform outputs from one state to another. This rules out activities which lower the probability of low income states but admits activities which raise output in low income states at the expense of output in high income states. Each individual has a production possibilities frontier (*PPF*) defined by $f(Y_1, \ldots, Y_S)=0$. It is assumed that the production set is convex.

The production decision is made simultaneously with the consumption decision by means of a market in contingent contracts which meets before productive resources are allocated. Individual equilibrium in this market results in prospective production and consumption as shown for the two-state case in Figure 6. Here *P* is the production point and *C* is the consumption point. In case of outcome one, the individual gives up income *OP* out of his production; in case of outcome two, he receives income *OC* in addition to his production. The equilibrium conditions are the familiar ones: at the price ratio indicated by the slope of line *m*, the points *P* and *C* are dictated by the necessity that the *PPF* and some indifference curve are tangent to one and the same line with slope *m*.

News changes the indifference maps of individuals. If it arrives before the market opens, it causes a shift of production and consumption in the direction of the state the news favors. News arriving after resources are irrevocably allocated has no effect, for this is the pure exchange case. The intermediate case, when the market has reached equilibrium but resources can be reallocated, is the most realistic.[21] In this case the new production point is shown as *P′* in Figure 6. The shift from *P* to *P′* is caused by news that outcome one is more probable than was previously thought. The production point being determined, the assets that the individual can trade are found by adding his prospective production for each state to the assets he has in that state; if he has debts, these must be subtracted from what he produces. The individual trades from the point *Z*, which is the vector sum of the production *P′* and his previous portfolio $\overrightarrow{PC}$. In this case he gains from reallocation of productive resources, but he loses on his portfolio.



P	initial production point
C	initial consumption point
P′	production point after receipt of news
Z	effective endowment after receipt of news
m	price line

FIGURE 6. TRADE IN CONTRACTS WITH PRODUCTION

B. *Public Information*

The value of public information depends upon when the news arrives. If it arrives after the contingent contracts market has met and after productive resources have already been allocated, it is neither useful nor harmful. Consider the case in which the news arrives before the allocation of productive resources and before the meeting of the contingent claims market. As in the pure exchange economy, the news has a redistributive impact. Individuals who are comparatively productive in a certain state are hurt if that state is less likely to occur and helped if that state is more likely to occur. Unless everyone is identical in risk preference and productive capacity, information involves a gamble which is detrimental to *ex ante* satisfaction.

As in the case of pure exchange, the value of information cannot be unambiguously assessed in the absence of efficient hedges in a prior market. Unlike the pure exchange case, this market is not efficient if it contains only *S* different contracts. It is necessary to introduce new contracts because the information has an impact on production as well as consumption. The natural extension in this instance is to a market in $2S$ different contracts, one for each possible state and each condition. The typical individual emerges from this market with a portfolio consisting of $Y_1^A, \ldots, Y_S^A$; $Y_1^B, \ldots, Y_S^B$, with probabilities $p_A\, {}_A\pi_1, \ldots, p_A\, {}_A\pi_S$; $p_B\, {}_B\pi_1, \ldots, p_B\, {}_B\pi_S$; and a corresponding set of intended outputs $\bar{Y}_1^A, \ldots, \bar{Y}_S^A$; $\bar{Y}_1^B, \ldots, \bar{Y}_S^B$. When news arrives showing which condition is applicable, productive resources are allocated accordingly. As usual, market equilibrium represents an efficient allocation of resources *ex ante.*[22]

---

[21] See Hirshleifer (1971, p. 568).

The value of information is generally positive in this situation, for if information is not purchased there can be no difference between Y_s^A and Y_s^B, which represents a constraint in the model. The exceptions to this are the pure exchange economy, already discussed, which can be represented in this context as meaning that every production possibility set is rectangular, and certain other examples in which kinks in the *PPF* of individuals prevent the adjustment of production between conditions A and B. Excepting these circumstances and given that impacts are insured, the social value of public information is generally positive. Whether this value is greater or less than the cost has to be decided in each individual case.

The above derivation imposes a condition for optimization of the value of information which may not be met in practice There is no assurance that the financial markets will have the variety of assets necessary to replicate the market in $2S$ contracts. There is an incentive to introduce new contracts to insure the impacts of information, but what of the greater number of contracts needed? This problem is reduced by noting that the natural extension of the market to $2S$ contracts is not necessary for efficient allocation. It can be replaced by a hedging market in two contracts which meets prior to the receipt of information, coupled with a meeting of the market in S contracts after the information has arrived. It is assumed that at the time trade is opened in two contracts, prior to the receipt of information, individuals correctly forecast the absolute prices that will prevail in equilibrium in the market in S contracts under conditions A and B.

The objects traded in this special hedging market may be thought of as contingent purchasing power for conditions A and B, respectively. The "endowments" of traders in this market are the values of their production in conditions A and B, and the "utility" functions of contingent purchasing power are derived from the utility functions for contingent contracts in conditions A and B.[23] The possibility

---

[22] Formally, each individual will maximize $p_A\sum_{s=1}^{S} {}_A\pi_s U(Y_s^A)+p_B\sum_{s=1}^{S} {}_B\pi_s U(Y_s^B)$ subject to $\sum_{s=1}^{S} P_s^A(\bar{Y}_s^A-Y_s^A)+\sum_{s=1}^{S} P_s^B(\bar{Y}_s^B-Y_s^B)=0$, $f(\bar{Y}_1^A,\ldots,\bar{Y}_S^A)=0$, and $f(\bar{Y}_1^B,\ldots,\bar{Y}_S^B)=0$, where the last two conditions indicate that production is on the production possibility frontier. Conditions of market equilibrium are that

$$\sum_{j=1}^{n} \bar{Y}_s^{Aj} = \sum_{j=1}^{n} Y_s^{Aj} \qquad s = 1, \ldots, S$$

$$\sum_{j=1}^{n} \bar{Y}_s^{Bj} = \sum_{j=1}^{n} Y_s^{Bj} \qquad s = 1, \ldots, S$$

[23] Individuals maximize $p_A U(I^A|A)+p_B U(I^B|B)$ subject to $f(\bar{Y}_1^A,\ldots,\bar{Y}_S^A)=0$, $f(\bar{Y}_1^B,\ldots,Y_S^B)=0$, and $Q_A(\sum_s P_s^A\bar{Y}_s^A-I^A)+Q_B(\sum_s P_s^B\bar{Y}_s^B-I^B)=0$. Here Q_A and Q_B are prices and the meaning of $U(I^A|A)$ and $U(I^B|B)$ is derived from the predicted equilibrium prices, P_s^A and P_s^B by way of the subsidiary individual maximum problems: $U(I^A|A)=\max. \sum_s {}_A\pi_s U(Y_s^A)$ subject to $\sum_s P_s^A Y_s^A - I_s^A = 0$, and similarly for $U(I^B|B)$. Equilibrium in the prior market requires $\sum_j I^{Ai} = \sum_j \sum_s P_s^A \bar{Y}_s^{Ai}$ and $\sum_j I^{Bi} = \sum_j \sum_s P_s^B \bar{Y}_s^{Bi}$. The prices P_s^A and P_s^B are equilibrium prices of the conditional markets in the sense that they induce $\sum_j \bar{Y}_s^{Ai} = \sum_j Y_s^{Ai}$ for all s and $\sum_j \bar{Y}_s^{Bi} = \sum_j Y_s^{Bi}$ for all s.

Proof that this is equivalent to the market described in fn. 22 is straightforward. If $P_s^A, P_s^B, \bar{Y}_s^A, Y_s^A, \bar{Y}_s^B, Y_s^B$ represent an equilibrium in the system of $2S$ markets, this becomes an equilibrium of the system of $S+2$ markets with addition of $Q_A=Q_B=1$, $I^{Ai}=\sum_s P_s^A Y_s^{Ai}$, and $I^{Bi}=\sum_s P_s^B Y_s^{Bi}$. Conversely, if $P_s^A, P_s^B, Y_s^A, Y_s^A, Y_s^B, Y_s^B$, of this reduction from $2S$ contracts to $S+2$ contracts makes it more likely that the possible impacts of research and information gathering will be properly insured, and less likely that important risks are being inflicted on individuals by the pursuit of public information.[24]

C. *A Note on Disagreement Concerning Probabilities*

The pure exchange model with disagreement concerning probabilities is similar to the production and exchange model just described. In order to insure the impacts of information, markets in $2S$ (or $S+2$) contingent contracts must be held, and *public information may have positive social value*. A single example is sufficient to illustrate this and prove that public information is not necessarily valueless when there is disagreement concerning probabilities.

Assume there are two traders and two possible outcomes,[25] that the traders' initial endowments are skewed, and that condition A favors trader number one. The probabilities ${}_A\pi_s$ and ${}_B\pi_s$ involved in conditions A and B are agreed upon by the two traders, but they disagree about the probabilities p_A and p_B of the conditions A and B. Trader one thinks condition A, which favors him, is very likely, while trader two thinks it is very unlikely. These traders can agree to purchase information and both can benefit in terms of *ex ante* satisfaction.[26] It is not evident whether examples can be found in which public information has negative social value, in spite of its impacts being fully insured by efficient trade in $2S$ (or $S+2$) contracts prior to the receipt of news.

D. *Private Information*

The incentives to obtain private information in a production and exchange economy, to take a position based on the news, and then to release the news are the same as in the pure exchange case. Private information is valuable to the individual, of course, by much the same argument as used before. Consider Figure 7 for purposes of exposition. In the absence of information the individual has indifference surfaces Ω°, the market price line has slope m, and consumption is at C°. If condition A obtains, the arrival of information shifts his indifference contour to Ω and consumption to C without affecting the production point, which depends only on the slope of m. In case of condition B, the consumption point is on the budget line to the northwest of C°. Section Ic gave a formal proof that *ex ante* satisfaction is higher if the information is purchased in this situation.

Further gains can be made by announcing the information. When the market is reopened, given that condition A obtains,

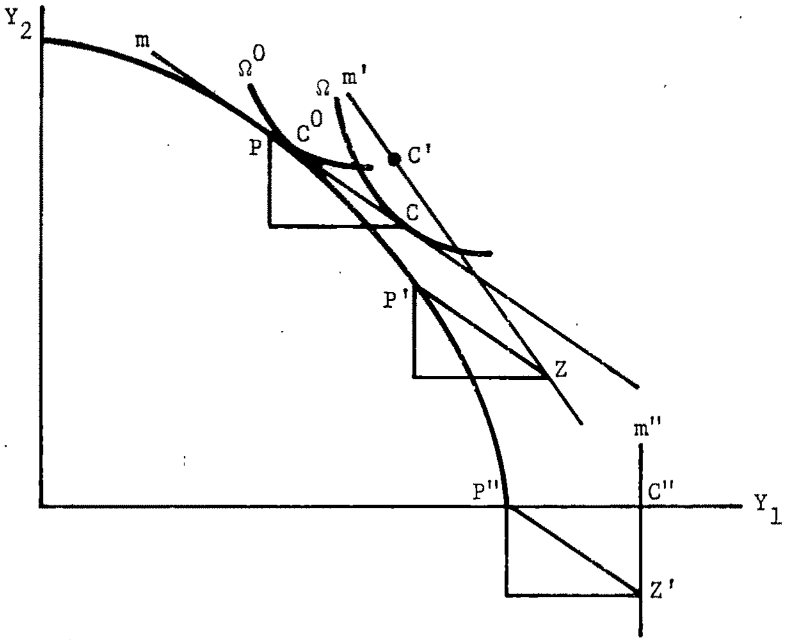
---

I^A, I^B, Q_A, Q_B is an equilibrium of the system of $S+2$ markets, it is equivalent to another equilibrium of this system where $P_S^{A\prime}=P_S^A(Q_A/Q_B)$, $I^{A\prime}=I^A(Q_A/Q_B)$, and $Q_A{}'=Q_B$. The price system $P_S^{A\prime}$, P_S^B and the quantities associated with the equilibrium are equilibrium prices and quantities in the system of $2S$ markets. The simplified structure of the model assures that second-order conditions are satisfied wherever first-order conditions are.

[24] Arrow emphasizes the possibility of a similar reduction in the number of contingent contracts required. That result is analytically very similar to the conclusion of this section, although the interpretation there involves different states rather than different conditions.

[25] The numerical example assumes $\overline{Y}_1^1=.9$, $\overline{Y}_1^1=.1$, $\overline{Y}_1^2=.1$, $\overline{Y}_2^2=.9$, ${}_A\pi_1=.75$, ${}_B\pi_1=.25$ (traders agree on this), but $p_A^1=.9$ (probability of condition A as seen by trader one), and $p_A^2=.1$ (probability of condition A as seen by trader two). Utility functions are $U(Y)=ln\ Y$. If they trade without information, equilibrium has $Y_1^1=.7$, $Y_2^1=.3$, $Y_1^2=.3$, $Y_2^2=.7$ and the price ratio is unity. The utility of equilibrium is $-.611$ for each individual. If they stand pat at the endowment point and purchase the information, the conditional posterior incomes are $Y_1^{A1}=.7$, $Y_2^{A1}=.7$, $Y_1^{A2}=.3$, $Y_2^{A2}=.3$, $Y_1^{B1}=.3$, $Y_2^{B1}=.3$, $Y_1^{B2}=.7$, $Y_2^{B2}=.7$. The utility of this arrangement is $-.441$ for each individual. The purchase of public information is socially valuable, even though in this example the impacts were less than perfectly insured.

[26] Contrast Hirshleifer (1971, p. 569) who maintains that in the example given here the purchase of public information is socially wasteful.



Ω^0	initial indifference curve
m	initial price line
P	initial production point
C^0	initial consumption point
C	consumption after receiving information
Ω	indifference curve after receiving information
m'	price line after announcing information
P'	production point after announcing information
C'	consumption after announcing information
Z	net assets after announcing information and shifting production
P'',C'',Z'	production, consumption, and assets when the news makes outcome one perfectly certain
m''	price line if outcome one is perfectly certain

FIGURE 7. ANNOUNCEMENT AND SPECULATION

the new equilibrium price ratio m' will lead to the production point P'. Producing at P' and holding the portfolio of assets and debts $\overrightarrow{PC}$, the individual trades from the point Z, which is the vector sum of the two. This permits him to reach consumption C', which is definitely better than C. Thus, private information is always made public, given the assumptions of the model and provided that the special nature of the information does not prevent announcement. As before, the case for announcement is even stronger when there is conscious speculation, for the individual can do better than C' by taking a position southeast of C on m before making his information known.[27]

---

[27] Hirshleifer assumes that the holder of private information will not release it, and argues that he will not reallocate investments either, because the market price of outputs is unvarying (1971, p. 566). This slip contributes to the strength of his conclusion that for a pro-

Each single piece of induced public information involves three costs: the purchase cost, redistributive cost, and the cost of increased risk bearing. In spite of the costs associated with it, there is no reason that induced public information has to be a waste of resources. If the redistributive effect is acceptable, if the impacts of news are adequately insured, and if the possibilities for transformation between states are favorable (i.e., if the *PPF* does not have corners), then losses due to the gamble are slight and the gain from adjusting production is large. This is less efficient than the purchase of public information with full insurance of its impacts, but it may be better than no information at all.

Considering all opportunities to purchase private information, one can ask whether the incentives to purchase it are more or less than required for efficiency. This question has a precise answer only under stringent assumptions. Assume that the producer of information receives a "superpatent" by which he extracts the entire value of the information, leaving everyone except himself no better off than before. If he also gains by private speculation the total incentive is certainly greater than the social value.[28] In the absence of superpatents, the question of adequate incentives must be asked for each project individually. If potential speculative gains from the project are great, they may exceed the unappropriated social value of the information; if they are small, the reverse can be true. Since speculative gain depends on secrecy, disguise, and ability to announce—which vary among persons as well as among kinds of information—the outcome depends on details that differ from case to case.

Individuals rank opportunities to spend on information according to their private value. Since potential speculative gains differ from project to project this ranking does not agree with a ranking according to social value. When private investors push expenditures to the margin of private profitability, they undertake some socially wasteful projects and pass over some that are socially valuable. Even if private purchasers together spend the socially appropriate amount of resources, which can happen only by accident, they still buy the wrong information.

It must be recalled that the discussion here deals with information made public by the sequence of speculation and announcement. Induced public information is never as valuable socially as the same information purchased collectively after fully insuring its impacts. In that sense the above comparison of social and private rankings of information-gathering projects is relevant only when constraints rule out the optimum mechanism of purchasing public information.

REFERENCES

K. J. Arrow, "Le rôle des valeurs boursières pour la répartition la meilleure des risques," *Econométrie*, CNRS, Paris, 1953, 41–7; reprinted "The Role of Securities in the Optimal Allocation of Risk Bearing," *Rev. Econ. Stud.*, Apr. 1964, *51*, 91–96.

W. Brainard and F. T. Dolbear, "Social Risk and Financial Markets," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 360–70.

G. Debreu, *Theory of Value*, New York 1959.

J. Drèze, "Market Allocation Under Uncertainty," *Eur. Econ. Rev.*, Winter 1970–71, *2*, 133–65.

---

duction and exchange economy, "the incentives for the use of resources to generate private information remain excessive" (p. 567). The further discussion quite rightly drops the insistence upon "excessive incentives" in the production and exchange economy, and the final conclusions are not really affected. If the news makes state one perfectly certain, as Hirshleifer assumed, the result is illustrated by the points P'', C'', Z', and m'' in Figure 7.

[28] This clarification was made by Hirshleifer in correspondence.

E. F. Fama and A. B. Laffer, "Information and Capital Markets," *J. Bus., Univ. Chicago*, July 1971, *44*, 289–98.

J. Hirshleifer, "The Private and Social Value of Information and the Reward to Inventive Activity," *Amer. Econ. Rev.*, Sept. 1971, *61*, 561–73.

———, "Foundations of the Theory of Speculation," working pap. 189, West. Man. Sci. Inst., June 1972.

R. Kihlstrom and M. Pauly, "The Role of Insurance in the Allocation of Risk," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 371–79.

I. H. Lavalle, "On Cash Equivalents and Information Evaluation in Decisions Under Uncertainty," *J. Amer. Statist. Assn.*, Mar. 1968, *63*, 252–90.

B. Malkiel, review of Manne, *J. Bus., Univ. Chicago*, Apr. 1968, *41*, 263–65.

H. Manne, *Insider Trading and the Stock Market*, New York 1966.

P. A. Samuelson, "Probability, Utility, and the Independence Axiom," *Econometrica*, Oct. 1952, *20*, 670–78.

L. J. Savage, *The Foundations of Statistics*, New York 1954.

Testing for the Existence of a Consistent Aggregate Index of Labor Inputs

By ERNST R. BERNDT AND LAURITS R. CHRISTENSEN*

Most empirical studies of production relationships have been based on aggregate indexes of capital and labor inputs. The aggregation of diverse capital inputs has drawn more criticism, most notably from the Cambridge school, than the aggregation of diverse labor inputs. It is often argued that labor inputs are highly substitutable one for another, and therefore that little is lost by combining them into an aggregate labor index. It is of course true that if the labor inputs are perfectly substitutable, nothing is lost by aggregation. It does not follow, however, that little is lost when labor inputs are highly but imperfectly substitutable. We have recently shown (1973a) that except for the case of zero or perfect substitutability, a consistent aggregate of diverse inputs exists if and only if certain Allen partial elasticities of substitution (*AES*, denoted σ_{ij}) are equal; the equality restrictions for consistent aggregation of labor inputs do not involve the *AES* between pairs of labor inputs but rather between individual types of labor inputs and nonlabor inputs.[1]

In this paper we estimate a production function for *U.S.* manufacturing with labor input disaggregated, and then test whether a consistent aggregate exists. It would be desirable to use data on hours worked by employees classified by personal characteristics such as age, sex, race, education, experience, and health; but such data are extremely difficult to construct. The U.S. Bureau of Labor Statistics does, however, classify employees in *U.S.* manufacturing on the basis of occupation. The category of nonproduction workers includes occupations which in general require more skill and training than the occupations included in the category of production workers. This split in the manufacturing work force allows us to estimate a production function with two labor inputs and one capital input. We denote production workers by *B* for blue collar, nonproduction workers by *W* for white collar, and capital by *K*.

Finis Welch and Zvi Griliches (1969, 1970) have hypothesized that capital is more complementary with skilled labor than with unskilled labor. Welch (pp. 36–38) suggests this as a partial explanation for the evidence that rates of return to education have apparently failed to decline with large increases in average education levels. Griliches (1970, pp. 104–05) employs the complementarity hypothesis as a partial explanation for the related evidence that skilled and unskilled income differentials have remained stable in spite of large increases in the supply of skilled labor. Griliches (1969) and Alvin Cook present some preliminary empirical evidence supporting the complementarity hypothesis, but neither study provides estimates of the relevant *AES*.

---

* Assistant professor of economics, University of British Columbia, and associate professor of economics, University of Wisconsin, respectively. This paper was written while we were at the Mathematics Computation Laboratory, Office of Emergency Preparedness, and the Price Research Division, U.S. Bureau of Labor Statistics, respectively, in Washington, D.C. Orley M. Ashenfelter, W. Erwin Diewert, Arthur S. Goldberger, Zvi Griliches, Dale W. Jorgenson, Lawrence J. Lau, and Charles E. Metcalf provided helpful comments on previous drafts, but we retain responsibility for any remaining errors. Christensen acknowledges research support from the National Science Foundation and the University of Wisconsin Graduate School.

[1] We rule out the unlikely situation where prices or quantities always move in fixed proportions due only to market conditions.

In an international cross-section study, Samuel Bowles has assumed that distinct labor inputs (differentiated by levels of schooling) are weakly separable from capital input, i.e., that a consistent aggregate index of these distinct labor inputs exists. This separability assumption implies that the *AES* between capital and one type of labor is equal to the *AES* between capital and any other type of labor.

For our application the Welch-Griliches complementarity hypothesis is $\sigma_{BK}>\sigma_{WK}$. The Bowles separability hypothesis is $\sigma_{BK}=\sigma_{WK}$. These hypotheses cannot both be true. To test these competing hypotheses, we require a production function in which the *AES* are not constrained a priori. We use the translog production function recently proposed by Laurits Christensen, Dale Jorgenson, and Lawrence Lau (1971, 1973).[2] We follow the empirical implementation for three inputs developed by the authors (1973b).

Our empirical results are clearcut: We decisively reject the hypothesis that $\sigma_{BK}=\sigma_{WK}$ and instead confirm the Welch-Griliches hypothesis that $\sigma_{BK}>\sigma_{WK}$. For the translog function the *AES* are not constant; they vary with the levels of the inputs. During the period 1950-1968, the average *AES* estimates are $\sigma_{BK}=3.72$, $\sigma_{WK}=-3.77$, and $\sigma_{BW}=7.88$. We conclude that even though production and nonproduction workers in *U.S.* manufacturing are highly substitutable, there is no way to consistently aggregate them. We further conclude that nonproduction workers and capital are complements, while production workers and capital are substitutes.

I. The Three-Input Translog Function for Production Labor, Nonproduction Labor, and Capital[3]

We assume that in the overall production function for *U.S.* manufacturing there exists an input function $F(B,W,K)$ which is weakly separable from all other inputs and exhibits constant returns to scale (*CRTS*). Any technical change affecting B, W, and K is assumed to be Hicks-neutral. These assumptions are sufficient to allow us to analyze substitution possibilities among B, W, and K without regard to the levels of other inputs or the level of technology. Further, we assume that the input function F can be represented by the translog form

$$\begin{aligned} \ln F = {} & \ln \alpha_0 + \alpha_B \ln B + \alpha_W \ln W + \alpha_K \ln K \\ & + 1/2\, \gamma_{BB}(\ln B)^2 + \gamma_{BW} \ln B \ln W \\ & + \gamma_{BK} \ln B \ln K \\ & + 1/2\, \gamma_{WW}(\ln W)^2 + \gamma_{WK} \ln W \ln K \\ & + 1/2\, \gamma_{KK}(\ln K)^2 \end{aligned}$$

Assuming that markets are competitive, the necessary conditions for efficient production are $\partial F/\partial B=P_B$, $\partial F/\partial W=P_W$, and $\partial F/\partial K=P_K$ where P_B, P_W, and P_K are input prices relative to the price of aggregate input F. For B this implies the equivalent condition

$$\frac{\partial \ln F}{\partial \ln B} = \frac{\partial F}{\partial B}\cdot\frac{B}{F} = \frac{P_B B}{F} = M_B$$

where M_B is the cost share of B in the total cost of producing F, i.e., $M_B=P_BB/(P_BB+P_WW+P_KK)$. Taking the logarithmic partial derivatives $\partial \ln F/\partial \ln B$, $\partial \ln F/\partial \ln W$, $\partial \ln F/\partial \ln K$, and equating them to the cost shares, we find

---

[2] Independently, the translog form has been discussed by Griliches-Ringstad and Sargan as a generalization of Kmenta's approximation to the *CES* production function. Chu, Aigner, and Frankel have also proposed a form which is quadratic in the logarithms. Christensen, Jorgenson, and Lau (1973) have estimated a two-input, two-output translog form using aggregate *U.S.* data.

[3] This section is based on our 1973b article, pp. 88–93, which develops the three-input empirical implementation in greater detail.

$$
\begin{aligned}
(1)\quad M_B &= \alpha_B + \gamma_{BB}\,ln\,B + \gamma_{BW}\,ln\,W \\
&\quad + \gamma_{BK}\,ln\,K \\
M_W &= \alpha_W + \gamma_{BW}\,ln\,B + \gamma_{WW}\,ln\,W \\
&\quad + \gamma_{WK}\,ln\,K \\
M_K &= \alpha_K + \gamma_{BK}\,ln\,B + \gamma_{WK}\,ln\,W \\
&\quad + \gamma_{KK}\,ln\,K
\end{aligned}
$$

We propose to estimate the parameters of the translog function F from these three semilogarithmic equations. We assume that the translog form is an exact representation of the input function F. Any deviations of the cost shares from the logarithmic marginal products are assumed to be the result of errors in optimizing behavior. Thus we specify a classical additive disturbance for each of the equations in (1). For applications to aggregate data the input price and quantity levels of the three inputs should be treated as endogenous variables; in this case there may be a problem of simultaneous equations bias in the parameter estimates by ordinary least squares. Therefore we estimate the parameters of each equation by two-stage least squares (*2SLS*). We denote the estimates of the α_i by a_i and the estimates of the γ_{ij} by c_{ij}. With this estimation method we obtain two distinct estimates of the three parameters which appear in two equations—γ_{BW}, γ_{BK}, and γ_{WK}. Let us distinguish the two estimates by a superscript representing the equation, e.g., we have c^B_{BW} and c^W_{BW} as the two estimates of γ_{BW}. Since the cost shares M_B, M_W, and M_K sum to unity at each observation, the *2SLS* parameter estimates must satisfy the following relations:

$$
\begin{aligned}
(2)\qquad a_B + a_W + a_K &= 1 \\
c_{BB} + \overset{W}{c}_{BW} + \overset{K}{c}_{BK} &= 0 \\
\overset{B}{c}_{BW} + c_{WW} + \overset{K}{c}_{WK} &= 0 \\
\overset{B}{c}_{BK} + \overset{W}{c}_{WK} + c_{KK} &= 0
\end{aligned}
$$

Of the twelve estimated parameters, only eight are free; the parameter estimates from any of the three equations can be derived from the parameter estimates of the other two equations. Furthermore, the parameter estimates are identical regardless of the two equations chosen for estimation. Henceforth we limit our consideration to estimation of equations M_W and M_K and derive estimates of the parameters in M_B from (2).

Equations (1) can be considered logarithmic marginal productivity relations of a well-defined production function if and only if their partial derivatives are symmetric in the inputs, for example, if γ_{BW} in M_B is equal to γ_{BW} in M_W. Estimates of the parameters, however, will not in general satisfy this requirement. Hence this aspect of the specification is subject to statistical confirmation.

We can impose symmetry by requiring that:

$$
(3)\qquad \overset{B}{c}_{BW} = \overset{W}{c}_{BW}, \qquad \overset{B}{c}_{BK} = \overset{K}{c}_{BK}, \qquad \text{and} \quad \overset{W}{c}_{WK} = \overset{K}{c}_{WK}
$$

Since c^B_{BW} and c^B_{BK} do not appear in our two estimating equations, we substitute (2) into (3) which yields the following equivalent set of cross equation symmetry restrictions:

$$
(4)\quad \overset{W}{c}_{WK} = \overset{K}{c}_{WK}, \qquad \overset{W}{c}_{BW} = -\,(c_{WW} + \overset{W}{c}_{WK}), \qquad \text{and} \quad \overset{K}{c}_{BK} = -\,(\overset{K}{c}_{WK} + c_{KK})
$$

Whether the loss of fit from imposing these three restrictions is significant can be assessed by standard statistical techniques as discussed below.

We note that the imposition of the column sum restrictions (2) and the symmetry restrictions (3) imply zero row sums. Zero row sums and zero column sums are necessary and sufficient for *CRTS*. Thus given (2), the imposition of symmetry results in the imposition of *CRTS*.

The five free parameters which we estimate directly are α_W, α_K, γ_{WW}, γ_{WK}, and γ_{KK}. The cross-equation symmetry restrictions cannot be imposed using equation by equation *2SLS*. One possible estimation procedure is to perform a constrained two-equation multivariate regression by minimizing the trace of the residual covariance matrix, i.e., estimating a "stacked" regression by *2SLS*. If the off-diagonal elements of the residual covariance matrix are nonzero, greater efficiency in estimation may be attained by taking the estimated covariance into account and using three-stage least squares (*3SLS*).[4] The problem with this approach is that the estimates of the translog parameters may no longer be independent of our choice of which two equations to estimate. The solution to this problem is to iterate the *3SLS* procedure until the weighted residual covariance matrix of the two equations converges to an identity matrix. The estimates of the five free regression coefficients obtained by this iterative three-stage least squares procedure (*I3SLS*) are independent of our choice of which two equations to estimate. Although Phoebus Dhrymes has shown that the *I3SLS* estimator is not in general the full information maximum likelihood estimator, the *I3SLS* estimator is asymptotically equivalent to the full information maximum likelihood estimator.

In our 1973a paper, we have established that weak separability restrictions are equivalent to certain equality restrictions on the Allen partial elasticities of substitution (*AES*). To illustrate, we note that for our three-input case the following are equivalent restrictions:

(i) B and W are functionally weakly separable from K, i.e., $F(B,W,K)=H[J(B,W),K]$;[5]

(ii) equality of the *AES* $\sigma_{BK}=\sigma_{WK}$;

(iii) the existence of a consistent aggregate price index P_L and a consistent aggregate quantity index L, with components P_B and P_W, B and W, respectively;

(iv) the path independence of a Divisia price index P_L and a Divisia quantity index L.[6]

With three inputs, three types of separability may exist: the separability of B and W from K (denoted BW-K), B and K from W (BK-W), and W and K from B (WK-B). The separability conditions on the three input translog functions are satisfied if and only if

$$\text{(5)}\quad BW\text{-}K:\quad M_B\gamma_{WK}-M_W\gamma_{BK}=0 \qquad (\sigma_{BK}=\sigma_{WK})$$

$$\text{(6)}\quad BK\text{-}W:\quad M_B\gamma_{WK}-M_K\gamma_{BW}=0 \qquad (\sigma_{BW}=\sigma_{WK})$$

$$\text{(7)}\quad WK\text{-}B:\quad M_W\gamma_{BK}-M_K\gamma_{BW}=0 \qquad (\sigma_{BW}=\sigma_{BK})$$

Only two of these three types of separability are independent, for if we subtract (5) from (6) we obtain (7). Alternatively, $\sigma_{BK}=\sigma_{WK}$ and $\sigma_{WK}=\sigma_{BW}$ implies $\sigma_{BK}=\sigma_{BW}$.

Substituting (1) into (5), (6), and (7), we obtain the conditions for global separability on the translog function:

$$\text{(8)}\quad BW\text{-}K:\quad \gamma_{BW}\gamma_{WK}-\gamma_{WW}\gamma_{BK}=0,\quad \alpha_B\gamma_{WK}-\alpha_W\gamma_{BK}=0$$

$$\text{(9)}\quad BK\text{-}W:\quad \gamma_{BK}\gamma_{WK}-\gamma_{KK}\gamma_{BW}=0,\quad \alpha_B\gamma_{WK}-\alpha_K\gamma_{BW}=0$$

$$\text{(10)}\quad WK\text{-}B:\quad \gamma_{WK}\gamma_{BK}-\gamma_{KK}\gamma_{BW}=0,\quad \alpha_W\gamma_{BK}-\alpha_K\gamma_{BW}=0$$

There is one set of linear restrictions on the five free parameters which satisfies each type of separability. They are presented in

---

[4] In the unconstrained *2SLS* regression, the regressors in the two equations are the same; taking the covariance into account would not alter these estimates.

[5] Since the partition under consideration is limited to two subsets, weak and strong separability are equivalent restrictions. This being the case, we hereafter denote weak or strong separability by the single word separability.

[6] See Hulten for further discussion.

TABLE 1—PARAMETER RESTRICTIONS FOR FUNCTIONAL SEPARABILITY

Separability Type	Linear Separability Conditions	Non-Linear Separability Conditions
BW-K	$\gamma_{KK}=0$ $\gamma_{WK}=0$ $(\sigma_{BK}=\sigma_{WK}=1)$	$\gamma_{KK}=\gamma_{WK}^2/\gamma_{WW}$ $\alpha_K=1+(\alpha_W\gamma_{WK}/\gamma_{WW})$ $(\sigma_{BK}=\sigma_{WK}\neq 1)$
BK-W	$\gamma_{WW}=0$ $\gamma_{WK}=0$ $(\sigma_{BW}=\sigma_{WK}=1)$	$\gamma_{KK}=\gamma_{WK}^2/\gamma_{WW}$ $\alpha_K=(\alpha_W-1)\gamma_{WK}/\gamma_{WW}$ $(\sigma_{BW}=\sigma_{WK}\neq 1)$
WK-B	$\gamma_{WW}=-\gamma_{WK}$ $\gamma_{WK}=-\gamma_{KK}$ $(\sigma_{BK}=\sigma_{BW}=1)$	$\gamma_{KK}=\gamma_{WK}^2/\gamma_{WW}$ $\alpha_K=\alpha_W\gamma_{WK}/\gamma_{WW}$ $(\sigma_{BK}=\sigma_{BW}\neq 1)$

the first column of Table 1. Each set of linear restrictions requires that a pair of *AES* equal one. For a pair of *AES* to be equal but not equal to one requires a set of non-linear restrictions to hold on the five free parameters. They are presented in the second column of Table 1. Of the six restrictions in (8), (9), and (10), only three are independent. The three independent restrictions can be mutually satisfied only if $\gamma_{BW}=\gamma_{BK}=\gamma_{WK}=0$. This reduces the translog function to the logarithmic Cobb-Douglas function with two free parameters, α_W and α_K. In this case all three kinds of separability hold simultaneously; we term this condition complete separability.

Our testing framework for the separability conditions is structured as follows. We first test whether the restrictions for complete global separability are satisfied. If these conditions are satisfied, we terminate our testing and conclude that all three types of separability are valid. On the other hand, if we reject this hypothesis we continue our testing procedure by investigating whether any one of the three sets of linear restrictions is satisfied. If we reject all three sets of linear restrictions, we continue our testing procedure by investigating whether any one of the non-linear separability conditions is accepted. If we reject all three sets of non-linear restrictions, we conclude that none of the separability conditions is satisfied.

We could test for complete global separability and the three types of linear separability by performing Bonferroni t-tests[7] on c_{BW}, c_{BK} and c_{WK}. Christensen (1973) has demonstrated, however, that the power of this test procedure is likely to be quite low. We choose the alternative procedure of assigning an overall significance level of .08, and allocating .01 to each of eight F-tests of joint hypotheses.

For each joint hypothesis we calculate the change in the weighted sum of squared residuals from imposing the restrictions divided by the number of restrictions. This ratio is then divided by the ratio of the weighted sum of squared residuals without the restrictions imposed to the number of residual degrees of freedom. The resulting test statistic is distributed, asymptotically, as F (ν_1, ν_2), where ν_1 is the number of restrictions and ν_2 is the number of residual degrees of freedom. Of course each F-statistic is distributed asymptotically as chi-squared divided by its number of degrees of freedom (restrictions imposed). These test statistics are asymptotically equivalent to likelihood ratio test statistics. The allocation of the overall signifi-

---

[7] See Miller, pp. 8, 15–16, 31–35, 67–70, and Christensen (1973) for a discussion of Bonferroni t-tests.

TABLE 2—SIGNIFICANCE LEVELS, RESTRICTIONS, RESIDUAL DEGREES OF FREEDOM, AND CRITICAL VALUES FOR TESTING THREE-INPUT TRANSLOG FUNCTION

Test	Significance Level	Restrictions	Residual Degrees of Freedom	Critical Values F	Critical Values χ^2/ν_1
Specification:					
Symmetry *CRTS*	.01	3	72	4.07	3.78
Separability:					
Complete Global Separability	.01	3	75	4.06	3.78
Three Linear Separability Tests	.01	2	75	4.90	4.60
Three Non-Linear Separability Tests	.01	2	75	4.90	4.60

cance level and the corresponding F and chi-square critical values are summarized in Table 2.

II. Data and Empirical Results

A. *Data*

The three inputs for this study are the services of production workers (B), nonproduction workers (W), and plant and equipment (K) in *U.S.* manufacturing. We have constructed the price and quantity indexes of K by Divisia aggregation of capital services from nonresidential structures and producer's durable equipment presented in our 1973b paper. The methodology follows Christensen and Jorgenson (1969).

Our construction of price and quantity indexes of labor services of production and nonproduction workers is based on unpublished data on employment, hours, and compensation from the U.S. Bureau of Labor Statistics and published data from the U.S. Office of Business Economics. Hours worked are adjusted for quality using the educational attainment index of Christensen and Jorgenson (1970) based on the development in Jorgenson and Griliches.

We treat the prices and quantities of the three inputs as jointly determined variables. The regressors in our estimating equations are logarithms of the quantity indexes of the three inputs. We use as instrumental variables for these regressors the fitted values from regressing them on a set of ten variables considered exogenous to the *U.S.* manufacturing sector. The ten exogenous variables are as follows: (1) *U.S.* population, (2) *U.S.* population of working age, (3) effective rate of sales and excise taxation, (4) effective rate of property taxation, (5) government purchases of durable goods, (6) government purchases of nondurable goods and services, (7) government purchases of labor services, (8) real exports of durable goods, (9) real exports of nondurable goods and services, and (10) *U.S.* tangible capital stock at the end of the previous year. These are the same exogenous variables used in our 1973b paper. The figures for 1929–68 are tabulated in the statistical appendix of that paper.

The three regressors, B, W, and K are scaled to equal 1.0 in 1949, the approximate midpoint for all three variables. We show in our 1973b paper that all our test results and elasticity of substitution estimates are independent of this scaling.

A detailed discussion of our data is presented in the Statistical Appendix.

B. *Empirical Results*

In the first column of Table 3 we present estimates of the unconstrained translog parameters based on *2SLS* regressions. The conventional R^2 figures (computed as one minus the ratio of the residual sum of squares to the total sum of squares in each equation) are .8999 for the B equation, .9056 for the W equation, and .1977 for the K equation. The Durbin-Watson statistics are 1.310 (B), .832 (W), and 2.994 (K). It appears that disturbances in the B and W equations are subject to a positive serial correlation.

We perform a test of the three symmetry restrictions, obtaining a test statistic of 8.90 indicating that we must reject the symmetry restrictions. However, the presence of serial correlation implies that some of the standard errors in Table 3 may be biased downward. Hence it is not clear whether the symmetry restrictions are valid. In the remainder of this paper we impose the symmetry restrictions as specified in (1); the rest of our test results are conditional on their validity.

TABLE 3—PARAMETER ESTIMATES OF THREE-INPUT TRANSLOG FUNCTION

Standard Errors in Parentheses

Parameter	Unconstrained	Symmetry Imposed
α_B	.6151	.6275
	(.0054)	(.0042)
γ_{BB}	.1782	.1833
	(.0128)	(.0129)
γ^{B}_{BW}	−.0768	−.1463
	(.0228)	(.0092)
γ^{B}_{BK}	−.0950	−.0370
	(.0238)	(.0097)
α_W	.2172	.2130
	(.0045)	(.0039)
γ^{W}_{BW}	−.1533	−.1463
	(.0107)	(.0092)
γ_{WW}	.0864	.1107
	(.0191)	(.0145)
γ^{W}_{WK}	.0666	.0357
	(.0200)	(.0123)
α_K	.1677	.1595
	(.0050)	(.0037)
γ^{K}_{BK}	−.0249	−.0370
	(.0117)	(.0097)
γ^{K}_{WK}	−.0078	.0357
	(.0210)	(.0123)
γ_{KK}	.0283	.0014
	(.0219)	(.0137)

In the second column of Table 3 we present estimates of the translog parameters with the symmetry restrictions imposed. The R^2 figures for these constrained estimates are .8284 (B), .8811 (W), and −.0233 (K).

James Durbin and Edmond Malinvaud have suggested that the conventional single equation Durbin-Watson statistic be used to check for serial correlation of disturbances in the multivariate and simultaneous equations setting. The appropriate number of degrees of freedom is (N^*,T), where T is the number of observations in each equation and N^* is the number of exogenous variables used in the first stage of the *I3SLS* estimation. For the current application $N^*=10$ and $T=40$. The Durbin-Watson statistics based on the symmetry-constrained regressions are .9374 (B), .7863 (W), and 2.351 (K). We test the null hypothesis of no autocorrelation against the alternative hypothesis of positive first-order autocorrelation. At a significance level of .05, the hypothesis of a zero first-order autocorrelation coefficient is rejected for the B and W equations, but is accepted for the K equation.

A further test on the translog specification is to assess whether the monotonicity and convexity conditions are satisfied for our estimated parameters. We check both estimated equations and the implied third equation to see if the predicted cost shares are positive. We find that they are positive at each observation; thus the monotonicity conditions are satisfied.

The isoquants of our estimated translog function are strictly convex if the bordered Hessian is negative definite. We find that the border Hessian is negative definite except for the years 1942–1944.[8]

---

[8] See our 1973b paper for further discussion of the computation of convexity conditions and *AES* for the translog function.

TABLE 4—ESTIMATED ALLEN PARTIAL AND SHADOW ELASTICITIES OF SUBSTITUTION: THREE-INPUT TRANSLOG FUNCTION WITH SYMMETRY IMPOSED

	1929	1939	1949	1959	1968	Mean 1950–68
AES						
σ_{BW}	6.31	7.00	10.03	5.73	5.51	7.88
σ_{BK}	3.14	3.35	4.32	2.98	2.92	3.72
σ_{WK}	−2.86	−3.62	−6.84	−2.20	−1.94	−3.77
σ_{BB}	−3.80	−3.85	−4.50	−3.86	−3.94	−4.16
σ_{WW}	−12.04	−14.44	−24.41	−9.91	−9.07	−14.82
σ_{KK}	−6.11	−6.49	−7.87	−5.74	−5.58	−6.40
SES						
S_{BW}	5.05	5.55	7.79	4.62	4.48	5.72
S_{BK}	2.15	2.23	2.67	2.08	2.07	2.30
S_{WK}	1.28	1.36	1.70	1.22	1.19	1.38

Our translog specification with symmetry imposed satisfies the monotonicity requirements for all years. The implied isoquants are strictly convex for all years except the World War II years 1942–1944. We conclude that the translog specification is attractive for assessing substitution possibilities in *U.S.* manufacturing.[9]

To measure factor substitution possibilities with our three-input translog function, we compute the Allen partial elasticities of substitution (*AES*) and the shadow elasticities of substitution (*SES*).[10] In Table 4 we report the *AES* (σ_{ij}) and *SES* (S_{ij}) for selected years and the 1950–68 arithmetic means. Analyzing the *AES* we find the striking result that $\sigma_{WK}<0<\sigma_{BK}$. Not only does K appear to be more complementary with W than with B as hypothesized by Welch and Griliches, but K and W appear to be complements by the usual Allen definition ($\sigma_{WK}<0$). Furthermore, we find that B is more substitutable with W than with K, $\sigma_{BW}>\sigma_{BK}$. The own *AES* are negative, as implied by convexity of the isoquants.

Turning to the *SES* we find that all estimates are positive as required by convexity. The same rankings of substitution possibilities are revealed by the *SES*, $S_{BW}>S_{BK}>S_{WK}$, as by the *AES*. Both the *AES* and the *SES* indicate that the *BW* and *BK* substitution possibilities have declined over time, whereas both σ_{WK} and S_{WK} have declined in absolute value. The estimates of *SES*, however, obscure the fact that W and K are complements.

Once we have computed the *AES* it is a simple matter to calculate the implied price elasticities of the three inputs. Allen, pp. 504–08, has shown that $E_{ij}=\partial ln\ X_i/\partial ln\ P_j=M_j\sigma_{ij}$. Thus we multiply the *AES* by the appropriate fitted cost shares and present the results in Table 5.

We now proceed to test for the functional separability restrictions. We first test for complete global separability, which is equivalent to testing for the Cobb-Douglas conditions $\sigma_{BW}=\sigma_{BK}=\sigma_{WK}=1$. The joint hypothesis $\gamma_{BW}=\gamma_{BK}=\gamma_{WK}=0$

---

[9] In addition to the above specification tests we tested the regression coefficients for stability with a Chow test. The hypothesis of equality of the parameters in the two subperiods 1929–46 and 1947–68 was rejected (chi-square test statistic of 3.97; critical value 2.21 at .05 level of significance). As with the symmetry test, however, the presence of serial correlation makes this result ambiguous.

[10] We have chosen the *AES* because it corresponds with conventional comparative statics analyses, i.e., it is proportional to $\partial ln\ X_i/\partial ln\ P_j$, when output and all other prices are held fixed but all quantities are allowed to adjust to a new equilibrium position. For purposes of comparison, we also present estimates of the *SES*. The *SES* are a linear combination of the *AES*; Mundlak has shown that $S_{ij}=((-M_iM_j)/(M_i+M_j))(\sigma_{ii}-2\sigma_{ij}+\sigma_{jj})$ for all $i\neq j$. See our 1973b paper for details on the computation of the *AES*.

TABLE 5—ESTIMATED OWN AND CROSS-PRICE ELASTICITIES THREE-INPUT TRANSLOG FUNCTION WITH SYMMETRY IMPOSED

Price Elasticity	1929	1939	1949	1959	1968	Mean 1950–68
E_{BB}	−2.166	−2.268	−2.826	−2.108	−2.104	−2.378
E_{WW}	−3.094	−3.509	−5.199	−2.735	−2.594	−3.570
E_{KK}	−1.057	−1.097	−1.251	−1.022	−1.004	−1.098
E_{BW}	1.622	1.701	2.136	1.581	1.576	1.789
E_{BK}	.543	.566	.687	.530	.526	.589
E_{WK}	− .495	− .443	−1.088	− .392	− .350	−.628
E_{WB}	3.597	4.123	6.299	3.129	2.942	4.198
E_{KB}	1.790	1.973	2.713	1.627	1.559	1.988
E_{KW}	− .735	− .880	−1.457	− .607	− .555	− .890

is decisively rejected; the test statistic is 96.23. Since complete global separability is rejected, we now test whether any one pair of separability restrictions is accepted. The test statistics for the linear and non-linear separability restrictions are presented in Table 6. All six pairs of separability restrictions are rejected at the .01 level of significance. We conclude that no consistent aggregate index of *BW*, *BK*, or *WK* exists for our *U.S.* manufacturing data. The "natural" aggregate of *B* and *W* has the lowest test statistic, but it is still quite decisively rejected. All the test statistics are more than twice the critical value; hence it is unlikely that our conclusions are affected by the persence of serial correlation.[11]

TABLE 6—TEST STATISTICS OF SEPARABILITY RESTRICTIONS

	Test Statistic
Linear Separability Restrictions	
BW-K: ($\sigma_{BK} = \sigma_{WK} = 1$)	10.76
BK-W: ($\sigma_{BW} = \sigma_{WK} = 1$)	134.19
WK-B: ($\sigma_{BW} = \sigma_{BK} = 1$)	144.33
Non-Linear Separability Restrictions	
BW-K: ($\sigma_{BK} = \sigma_{WK} \neq 1$)	144.27
BK-W: ($\sigma_{BW} = \sigma_{WK} \neq 1$)	37.81
WK-B: ($\sigma_{BW} = \sigma_{BK} \neq 1$)	22.75

III. Concluding Remarks

We have found that there is no way to construct a consistent aggregate index of blue and white collar workers in *U.S.* manufacturing. The vast majority of empirical studies on production behavior, however, have been based on relationships involving the aggregate indexes of capital and labor services. It is appropriate, therefore, that we examine the empirical significance of the separability (aggregation) assumption.

In his well-known paper "Technical Change and the Aggregate Production Function," Robert Solow has demonstrated that aggregate indexes of labor and capital are clearly satisfactory for explaining aggregate output. Of the five functional forms that Solow fitted, the lowest simple correlation coefficient between predicted and actual output was .9964.

Following Solow's lead, let us construct a predicted "output" series (actually aggregate input) F^* using the estimated translog parameters:

$$\ln F^* = a_B \ln B + a_W \ln W + a_K \ln K$$

With a Divisia index L of B and W, the conventional $\sigma_{KL} = S_{KL}$ averages approximately 1.42 for the two input translog function. See our 1973b paper for details.

[11] Although we conclude that no consistent aggregate of labor input exists, any number of procedures can be used to construct an inconsistent labor aggregate.

$$+ 1/2\, c_{BB}(ln\, B)^2$$
$$+ c_{BW}\, ln\, B\, ln\, W + c_{BK}\, ln\, B\, ln\, K$$
$$+ 1/2\, c_{WW}(ln\, W)^2$$
$$+ c_{WK}\, ln\, W\, ln\, K + 1/2\, c_{KK}(ln\, K)^2$$

We also construct a Divisia aggregate F of B, W, and K to serve as actual output. With only symmetry imposed on the translog parameters the simple correlation between F and F^* is .9999. Imposing the BW-K separability restrictions and computing F^*, we obtain a simple correlation with F of .9999 for the linear restrictions and .9994 for the non-linear restrictions. Since these restrictions are equivalent to using a labor aggregate, it is clear that we concur with Solow's finding. We conclude that if a researcher is only interested in explaining a time-series of output, he loses little by using aggregate inputs.

Suppose, however, that we are interested in explaining variations in factor shares. We compute the predicted factor shares, first with symmetry imposed, then with BW-K separability imposed. The correlations between these predicted shares and the actual shares are presented in Table 7.

TABLE 7—CORRELATIONS BETWEEN PREDICTED AND ACTUAL FACTOR SHARES

	Share		
Restrictions Imposed	*B*	*W*	*K*
Symmetry only	.9140	.9388	.2967
BW-K Linear Separability	.8742	.9032	.0000
BW-K Non-Linear Separability	.8639	.8347	.4394

The correlations between the predicted and actual shares are substantially lower for the B and W shares when the separability restrictions are imposed. The picture is mixed, however, with respect to the capital share. To help interpret this result we note that the BW-K linear separability restrictions require that the K share be constant, thereby forcing a zero correlation with the actual share. The sum of the B and W shares is also a constant. The non-linear separability restrictions, however, require the ratio of the B and W shares to be constant. This causes a significant decline in the fit because the trend in the B share is down while the trend in the W share is up. The result is an improvement in the fit for the K share.

We conclude that the cost of aggregation is relatively low if one wants only to explain output trends; but if one wants to explain trends in factor shares the cost of assuming aggregation (or, equivalently, separability) is significant.[12]

STATISTICAL APPENDIX

In a study published by the *Survey of Current Business*, F. Beatrice Coleman notes that both the Census Bureau and the Bureau of Labor Statistics (*BLS*) have defined the category of production workers as follows:

> "Production" workers include working foremen and all nonsupervisory workers (including leadmen and trainees) engaged in fabricating, processing, assembling, inspection, receiving, storage, handling, packing, warehousing, shipping, trucking, hauling, maintenance, repair, janitorial, watchmen services, product development, auxiliary production for plant's own use (e.g., power plant), and recordkeeping, and other services closely associated with the above production operations.
>
> Employees other than production workers include those engaged in the following activities: Executive, purchasing, finance, accounting, legal, personnel, cafeterias, medical, professional, and technical activities, sales, sales delivery (e.g., routemen), advertising, credit, col-

---

[12] Bowles has argued that little information is lost by aggregating labor inputs, because the predicted output of his labor subfunction is highly correlated with his actual output. The above discussion suggests that a more stringent measure of information loss by aggregation could be obtained by observing the correlation of the actual and predicted factor shares (logarithmic marginal products).

> lection, and in installation and servicing of own products, routine office function, factory supervision (above the working foreman level), and force account construction employees on payroll engaged in construction of major additions or alterations to the plant who are utilized as a separate work force. Those engaged in regular maintenance and repair operations are included with production workers. [p. 24]

The distinction between production and nonproduction workers is based on the degree of direct association with the physical production process. In a recent study, *BLS* (1969) has labeled nonproduction workers as "office and related workers" while production workers were labeled as "nonoffice workers." In the *OBE* study by Coleman, production labor was labeled as "wage labor" and nonproduction labor as "salaried labor." Note that with these alternative categorizations, all force account construction and cafeteria workers (labeled as nonproduction labor) are considered salaried or office and related workers. Nevertheless, the essential division of production, nonoffice, or wage labor from nonproduction, office and related labor, or salaried labor is a useful one; in that spirit we denote the quantity of production labor input as B (for blue collar) and the quantity of nonproduction labor input as W (for white collar).

A man-hours index is conventionally defined as the product of persons engaged and average hours per person. A man-hours series frequently used by economists is that of Kendrick (1961, 1972); his index for 1929–68 constructs man-hours of all labor input as the product of all persons engaged and average hours per production worker, normalized to 1958=100. Kendrick therefore implicitly assumes that average hours per nonproduction worker are equal to average hours per production worker.

Since 1947 the Office of Productivity and Technology, *BLS*, has taken samples to measure separately man-hours series for production and nonproduction labor in *U.S.* manufacturing. These underlying estimates are used to construct their published total man-hours series for all labor inputs. Let us denote MH = man-hours, P = production persons engaged, N = nonproduction persons engaged; then a total labor man-hours series is constructed as

$$\text{(A1)} \quad MH_t = (P/(P+N))MH_p + (N/(P+N))MH_n$$

where the subscripts t, p, and n refer to total, production, and nonproduction labor, respectively. *BLS* has published data series of N, P, average hours worked by production labor, and average pay per hour of production labor for the 1929–68 period.[13]

We construct the 1929–68 series on MH_t as follows: For the 1947–68 period, we use the *BLS* total man-hours series—a series which reflects changes in the average hours worked by production and nonproduction labor, as well as changes in N and P.[14] For the 1929–46 period, like Kendrick we assume that average hours of production labor are equal to average hours of nonproduction labor. In these years MH_t is computed as the product of persons engaged $(N+P)$ and average hours per production worker.

We construct the 1929–68 series on MH_p as follows: For the entire 1929–68 period we compute MH_p as the product of number of production workers and average hours per production worker. We then use the series on N, P, MH_t, and MH_p in (A1) to compute an implicit MH_n series for 1929–68.

The man-hours series of MH_p and MH_n do not reflect changes in the quality of labor. We incorporate quality adjustments by utilizing an index of educational attainment constructed by Christensen and Jorgenson (1970). Separate indexes of educational attainment for production and nonproduction labor are not available. The only published series of educational attainment in *U.S.* manufacturing is that of Griliches (1967), who constructed an educational attainment index of the male labor force 18 years and over for the years 1940, 1950, 1952, 1957, 1959, and 1962. Interpolating for missing years by assuming constant rates of change between benchmarks, we obtain an index of educational attainment for the male labor force 18 years and over in *U.S.* manufacturing 1940–62. Although the Christensen-Jorgenson index covers the entire *U.S.* labor force, we find that the Griliches index and the Christensen-Jorgenson index are highly correlated; the simple correlation coefficient between them is .9884. Since the Christensen-Jorgenson index spans the years 1929–68 and allows for changes in the educational attainment of females in the labor force, it was decided to retain the Christensen-Jorgenson index. Further, we assume that rates of educational attainment are the same for production and nonproduction labor.[15] The Christensen-Jorgenson educational index is multiplied by MH_p and MH_n; the resulting quantity series, labeled B for production labor and W for nonproduction labor, take account of changes in persons engaged, changes in average hours worked, and changes in the quality of labor as measured by changes in educational attainment.

---

[13] See *BLS* (1968). Nonproduction labor is calculated as total persons engaged minus production persons engaged. The 1968 figures are found in later issues of the *BLS* publication *Monthly Labor Review*.

[14] The Office of Productivity and Technology, *BLS*, kindly provided the underlying tables of their published total man-hours series.

BLS does not provide compensation series for total labor input in *U.S.* manufacturing prior to 1947. *OBE*, however, does provide such series. To reconcile the *OBE* and *BLS* series, the annual *OBE* series on full-time equivalent employees in *U.S.* manufacturing is distributed between production and nonproduction workers on the basis of their ratios in the *BLS* series. For the years 1947–68 the *OBE* and *BLS* series do not differ greatly; the slight differences that do exist may be attributed to distinctive sampling techniques. Further, in the *OBE* series the number of proprietors is calculated as the difference between persons engaged and full-time equivalent employees. We allocate proprietors to the category of nonproduction labor.

Separate payroll data for nonproduction and production labor are not available. We infer nonproduction payroll as the residual of total payroll minus production payroll.

To obtain a wage bill for production workers, we form the product of production workers, average hours per production worker, and average pay per hour for production labor. The wage bill for nonproduction labor is computed as the difference between total wages and salaries and the wage bill for production labor.

These wage bills for production and nonproduction labor do not represent the full labor costs borne by the employer, because supplements to wages and salaries are a significant part of total payroll. Separate time-series data on supplements to wages and salaries for production and nonproduction labor are not available. However *OBE* publishes a series on supplements to wages and salaries for all employees in *U.S.* manufacturing; it is the difference between compensation to employees and wages and salaries. These supplements to wages and salaries are allocated to production and nonproduction labor on the basis of their share of wages and salaries in total wages and salaries. In order to take account of the implicit compensation of proprietors, we assume that proprietors on the average earn the same compensation as the average compensation of all labor types in *U.S.* manufacturing.[16] We multiply average compensation (compensation to all employees divided by full-time equivalent employees) by the number of proprietors (persons engaged minus full-time equivalent employees), and allocate this compensation to nonproduction labor. Total compensation for production labor is the sum of production wages and salaries plus production supplements to wages and salaries. Total compensation for nonproduction labor is the sum of nonproduction wages and salaries plus nonproduction supplements to wages and salaries plus proprietors compensation.

A supplement to this Statistical Appendix is available on request from the Social Science Research Institute, 1180 Observatory Drive, University of Wisconsin, Madison, Wisconsin, 53706. It contains four tables of annual (1929–1968) figures for the following: 1) quantity indexes of the inputs B, W, and K; 2) antilogs of the instrumental variables for B, W, and K used in the *2SLS* and *I3SLS* regressions; 3) input costs in current dollars for B, W, and K; 4) estimated Allen and shadow elasticities of substitution.

---

[15] We are not aware of any study which has undertaken to estimate rates of educational attainment for production and nonproduction labor in *U.S.* manufacturing, in any *U.S.* industry, or for the entire *U.S.* labor force.

[16] See Christensen (1971) for a justification of this assumption.

REFERENCES

R. G. D. Allen, *Mathematical Analysis for Economists,* London 1938.

E. R. Berndt and L. R. Christensen, (1973a) "The Internal Structure of Functional Relationships: Separability, Substitution, and Aggregation," *Rev. Econ. Stud.,* July 1973, *40,* 403–10.

——— and ———, (1973b) "The Translog Function and the Substitution of Equipment, Structures, and Labor in U.S. Manufacturing 1929–68," *J. Econometrics,* Mar. 1973, *1,* 81–114.

S. S. Bowles, "Aggregation of Labor Inputs in the Economics of Growth and Planning: Experiments with a Two-Level CES Function," *J. Polit. Econ.,* Jan./Feb. 1970, *78,* 68–81.

G. C. Chow, "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica,* July 1960, *28,* 591–605.

L. R. Christensen, "Entrepreneurial Income: How Does it Measure Up?," *Amer. Econ. Rev.,* Sept. 1971, *61,* 575–85.

———, "Simultaneous Statistical Inference in the Normal Multiple Linear Regression Model," *J. Amer. Statist. Assn.,* June 1973, *68,* 457–61.

——— **and D. W. Jorgenson,** "The Measurement of U.S. Real Capital Input, 1929–1967," *Rev. Income and Wealth,* Dec. 1969, *15,* 293–320.

——— and ———, "U.S. Real Product and Real Factor Input, 1929–1967," *Rev. Income and Wealth,* Mar. 1970, *16,* 19–50.

———, ———, **and L. J. Lau,** "Conjugate Duality and the Transcendental Logarithmic Production Function," *Econometrica,* July 1971, *39,* 255–56.

———, ———, and ———, "Transcendental Logarithmic Production Frontiers," *Rev. Econ. Statist.,* Feb. 1973, *55,* 28–45.

S. F. Chu, D. J. Aigner, and M. Frankel, "On the Quadratic Law of Production," *Southern Econ. J.,* July 1970, *37,* 32–9.

F. B. Coleman, "Postwar Changes in Manufacturing Payrolls," *Surv. Curr. Bus.,* Dec. 1959, *35,* 19–24.

A. A. Cook, "The Demand for Cooperating Labor in Manufacturing Industries 1947–64," unpublished doctoral dissertation, Univ. Washington 1968.

P. J. Dhrymes, "Small Sample and Asymptotic Relations Between Maximum Likelihood and Three Stage Least Squares Estimators," *Econometrica,* Mar. 1973, *41,* 357–64.

J. Durbin, "Testing for Serial Correlation in Systems of Simultaneous Regression Equations," *Biometrika,* Dec. 1957, 370–77.

Z. Griliches, "Production Functions in Manufacturing: Some Preliminary Results," in M. Brown, ed., *The Theory and Empirical Analysis of Production,* New York 1967.

———, "Capital-Skill Complementarity," *Rev. Econ. Statist.,* Nov. 1969, *51,* 465–68.

———, "Notes on the Role of Education in Production Functions and Growth Accounting," in W. Lee Hansen, ed., *Education, Income, and Human Capital,* New York 1970, 71–115.

——— **and V. Ringstad,** *Economies of Scale and the Form of the Production Function,* Amsterdam 1971.

C. Hulten, "Divisia Index Numbers," *Econometrica,* forthcoming.

D. W. Jorgenson and Z. Griliches, "The Explanation of Productivity Change," *Rev. Econ. Stud.,* July 1967, *34,* 249–83.

J. W. Kendrick, *Productivity Trends in the United States,* New York 1961.

———, *Postwar Productivity Trends,* New York 1973.

J. Kmenta, "On Estimation of the CES Production Function," *Int. Econ. Rev.,* June 1967, *8,* 180–89.

E. Malinvaud, *Statistical Methods of Econometrics,* Chicago 1970.

R. G. Miller, *Simultaneous Statistical Inference,* New York 1966.

Y. Mundlak, "Elasticities of Substitution and the Theory of Derived Demand," *Rev. Econ. Stud.*, Apr. 1968, *35*, 225–36.

J. D. Sargan, "Production Functions," in P. R. G. Layard et al., eds., *Qualified Manpower and Economic Performance*, Part V, London 1971, 145–204.

R. Sato and T. Koizumi, "On the Various Concepts of Elasticity of Substitution and their Application in the Multifactor Production Process," mimeo., Brown Univ., Aug. 1970.

R. M. Solow, "Technical Change and the Aggregate Production Function," *Rev. Econ. Statist.*, Aug. 1957, *39*, 312–20.

F. Welch, "Education in Production," *J. Polit. Econ.*, Jan./Feb. 1970, *78*, 35–59.

U.S. Office of Business Economics, *The National Income and Product Accounts of the United States 1929–65: Statistical Tables*, Washington 1966.

U. S. Bureau of Labor Statistics, *Employment and Earnings Statistics for the United States 1909–1968*, Bull. 1312-6, Washington 1968.

———, *Employee Compensation in the Private Nonfarm Economy, 1966*, Bull. 1627, Washington 1969.

———, *Monthly Labor Review*, various issues.

The Stochastic Structure of the Velocity of Money

By John P. Gould and Charles R. Nelson*

Students of money and monetary policy have focused a substantial amount of attention on the behavior of the ratio of national income to the stock of money. In the most naive form of the quantity theory of money, this ratio—the velocity of money—is assumed to be a constant. Such an assumption enjoys no empirical support, however, because the velocity of money clearly changes over time. More modern and sophisticated versions of the quantity theory treat velocity as a stable function of other variables,

$$V_t = V(X_{1t}, X_{2t}, \ldots)$$

where the variables $X_{it}, \ldots$ include such things as the interest rates, changes in the price level, and the ratio of current to permanent income.[1] Indeed, much of the empirical money demand literature deals with estimation of money demand functions which incorporate at least a subset of these variables. To the extent that these efforts have been successful, velocity may be said to be predictable, at least in the sense of *ex post* predictability given the determinants of velocity. It has also been maintained that velocity is predictable in an extrapolation sense; that is, movements in velocity may to a degree be related to preceding history, allowing not only *ex post* explanation of a sort, but also forecasting. According to that view, "trends" and episodic deviations from trend may be discerned in the time-series of velocity. This paper studies the stochastic structure of velocity in order to determine whether there is a statistical basis for extrapolative prediction. It is important to emphasize that our statistical analysis focuses exclusively on the velocity series, since we are interested in knowing what a time-series analyst could conclude about velocity without reference to other economic variables.

We conclude that the velocity series constructed by Milton Friedman and Anna Schwartz is well characterized as a simple random walk. Another way of phrasing this conclusion is that whatever the determinants of velocity and their individual stochastic structures may be, their combined effect is such that successive changes in velocity are essentially uncorrelated. This would imply that of the past history available at any given date only the current observation is relevant for prediction. We also find that the apparent downward historical "drift" in velocity is not statistically significant. Further, the simple random walk model yields better postsample predictions than either a random walk-with-drift model or a constant-velocity model. Finally, we comment on the consistency of our findings with alternative fundamental views of the money-income relationship and with the kind of distributed lag equations featured in recent "monetarist" econometric models.

---

* Associate professors, Graduate School of Business, University of Chicago. We are grateful to Milton Friedman, Merton Miller, Harry Roberts, Stanley Fischer, William Poole, and members of the Money Workshop at the University of Chicago for their comments. Gould's participation in this research was supported by NSF Grant 2703 A#1 and Nelson's by NSF Grant 24501.

[1] For example, see Milton Friedman (1956).

I. The Friedman-Schwartz Velocity Series

The basis of the belief that velocity is a predictable quantity in the extrapolative sense appears to be in large part due to the relatively persistent decrease in this ratio over a long span of time. The Friedman and Schwartz velocity series declined from a peak of 4.97 in 1880 to a low of 1.16 in 1946.[2] Since the 1946 low, the series has generally increased and by 1960, according to the Friedman-Schwartz data, it had risen to 1.69. The decline from 1869 to 1946 was not monotonic, and sharp fluctuations have occurred during the postwar years. The postwar rise in velocity has caused no small concern to monetary theorists because it appears difficult to explain this behavior using arguments that are consistent with the experience of earlier years. As Friedman and Schwartz note in their discussion of the postwar behavior of velocity:

> The contrast with earlier experience is sharp. As far back as our figures go, the general tendency is for velocity of money to decline secularly at a rate that has averaged slightly more than 1 percent per year over the nine decades from 1869 to 1960. [p. 639]

They continue:

> Viewed in the long-term perspective provided by Chart 57, the postwar rise is the second of two surprisingly similar episodes both at variance with earlier behavior. The unprecedented collapse in the velocity of money from 1929 to 1932 is roughly matched by the corresponding collapse from 1942 to 1946; the rise from 1932 to 1942, interrupted from 1937 to 1940, is roughly matched by the rise from 1946 to 1960, interrupted from 1951 to 1954. As Chart 58 demonstrates, though the second movement has persisted for a somewhat longer period than the first, the two episodes are alike not only in general pattern and in the magnitude of the total rise and fall, but even in some of their minor movements. As befits episodes superimposed on a declining trend, the second movement occurs around a lower trend than the first. [pp. 639–41]

The interesting aspect of this discussion is that it appears to be based on the assumption that meaningful patterns in the movement of velocity and deviation from general trends can be inferred directly from an examination of the time-series of velocity. For example, the behavior of velocity in the two postwar "episodes" is analyzed in reference to the downward "trend" in velocity prior to the war. We find further allusions to such patterns as the authors continue:

> In this perspective, the postwar rise appears to be largely a reaction to the prior fall, just as the rise from 1932 to 1937 or 1942 appears to be a reaction to the fall from 1929 to 1932. But why did the velocity of money fall so sharply from 1929 to 1932 and again from 1942 to 1946, and why did the recovery take so long? In these terms, the puzzle is less why the velocity of money was "so high" in 1955 to 1960 than why it was so low and so variable during the preceding twenty-five years. [pp. 641–42]

Again, this discussion appears to draw inferences about perceived patterns and trends directly from the chart of the time-series of velocity. In particular, the term "reaction" in reference to rises in velocity relative to immediately preceding declines suggests dependence over time relating a given episode or deviation from trend to previous such movements.

The authors even appear to suggest that the patterns in velocity per se explain so much that the potential explanatory value of certain other variables is irrelevant. For example, in discussing the effects of interest rates on velocity in the postwar period they say:

> Even for the postwar period, however, this interpretation has a serious defect;

---

[2] The velocity series is tabulated in Table A5, p. 774.

> it explains too much. If interest rates account for the later postwar rise in the velocity of money, presumably they had a similar effect in the earlier postwar period. But the velocity rise in the earlier postwar years, as we have seen, can be accounted for as simply a recovery from abnormally low wartime years, so the interest rate explanation is superflous. [pp. 647–48]

II. A Stochastic Model for Velocity

Clearly, the assumption that velocity is stable or at least predictable is essential to any empirical interpretation of the *strict* monetarist position. It would seem natural to inquire, therefore, whether empirical evidence exists to support such an assumption. While few investigators would be satisfied with an assumption of *constant* velocity, the question of extrapolative predictability remains. Has the downward historical drift in velocity been significant in a statistical sense? Can secular movements in velocity be interpreted in any meaningful way as "trends" or "episodes" which may be linked to previous movements in velocity as Freidman and Schwartz have suggested? One approach to answering such questions is to examine the stochastic structure of the velocity series; in particular, its serial covariance structure. A tendency for velocity to move in reaction to preceding movements, for example, should be reflected in negative serial covariances or correlation at the appropriate lags. In any case, by associating stochastic structure with a statistical generating mechanism, we may be able to describe the behavior of the series and characterize its predictability in analytical terms.

The velocity series developed by Friedman and Schwartz is the longest consistent series available, spanning the years 1869 through 1960. Money is defined to include currency held by the public plus total deposits in commercial banks. Their money income numbers are based on Kuznets' net national income estimates, interpolated for 1956–60 from Department of Commerce estimates.[3] The data are plotted in Figure 1. Friedman has stated elsewhere (1968) that the data are least accurate for the early years through 1890. Consequently, in reporting results we shall distinguish between the full series and the shorter series comprising 1891–1960. From the plot of the data it is evident that measured velocity was subject to much greater fluctuation during the first sixteen years, through 1885, than during later years, whether because of actual fluctuations or because of the measurement errors to which Friedman refers.

Our initial presumption was that the velocity series, or its first differences, would have a representation as a member of the broad class of linear stochastic processes. This would include autoregressive, moving average, and mixed autoregressive-moving average (*ARMA*) processes. A general methodology for associating the sample autocorrelations of a time-series with particular *ARMA* models has been described by G. E. P. Box and G. M. Jenkins. A wide variety of patterns of autocorrelation may be shown to be implied by models belonging to this class. Sample autocorrelations of the long and short series and their first differences are presented in Table 1. Under the null hypothesis that the true autocorrelations are all zero, the individual standard errors of the sample autocorrelations are just $T^{-1/2}$ as indicated in the table.

The autocorrelations of both undifferenced velocity series are very large and decline slowly with increasing lag, a necessary result of the fact that the data cross the historical mean only once, in 1902, ex-

---

[3] See Friedman and Schwartz, Appendix A, fn. C.1, p. 775.

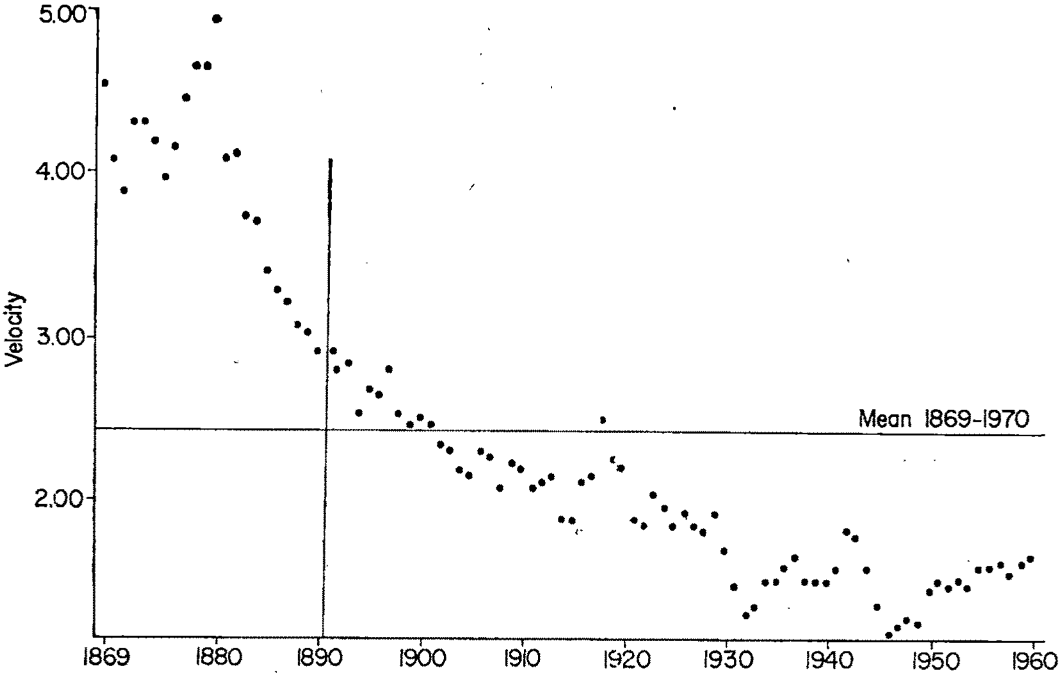


FIGURE 1

cept for a one-year "blip" in 1918.[4] Since this pattern is typical of a nonstationary series of the sort which displays no affinity for a mean value, the autocorrelations of the first differences of velocity are perhaps of greater interest. These tell a rather surprising story. Namely, only at lag three in the case of the full series do any of the autocorrelations approach anything like two standard errors in absolute magnitude. In general, the autocorrelations are small, exhibiting no pattern which might be suggestive of either autoregressive or moving average structure. In other words, a model which accounts for the observed series, denoted V_t, is the random walk

$$V_t - V_{t-1} = u_t + \delta \tag{1}$$

where $\{u_t\}$ is a sequence of zero mean uncorrelated random disturbances and δ is the expected value of the first differences $(V_t - V_{t-1})$. Another way of writing (1) expresses the current value of V_t as V_{t-1} plus the random disturbance u_t plus the constant δ, thus alternatively,

$$V_t = V_{t-1} + u_t + \delta \tag{2}$$

The evolution of the observed series may be described then as the accumulation of random changes plus a "drift" if δ is nonzero.[5]

---

[4] The formula for sample autocorrelations used in this study is

$$r_j = \frac{\sum_{t=1}^{T-j} (Z_t - \overline{Z})(Z_{t+j} - \overline{Z})}{\sum_{t=1}^{T} (Z_t - \overline{Z})^2}$$

where $\overline{Z}$ is the mean of the observations $Z_1, \ldots, Z_t$.

[5] It is important to note that the random walk hypothesis is also consistent with the observed sample autocorrelations of the *levels* of velocity. To see this note that model (2) is a special case of the Markov process $V_t = \phi V_{t-1} + u_t + \delta$, with $\phi = 1$. An approximation formula for the expectation of r_1 of a Markov process is given by Maurice Kendall and Alan Stuart, p. 435. Correcting their expression for number of terms in nu-

TABLE 1—SAMPLE AUTOCORRELATIONS OF FRIEDMAN-SCHWARTZ ANNUAL VELOCITY DATA, 1869–1960 AND 1891–1960

	1869–1960		1891–1960	
Lag	Levels	First Differences	Levels	First Differences
1	.95	−.08	.92	−.02
2	.92	.07	.84	−.07
3	.89	−.20	.76	−.18
4	.85	.04	.72	−.17
5	.81	−.06	.69	−.08
6	.78	.09	.66	.17
7	.74	−.03	.61	.05
8	.70	−.03	.56	−.02
9	.66	−.11	.53	−.01
10	.61	−.03	.48	.01
11	.56	.15	.44	.04
12	.50	−.13	.40	−.12
13	.46	.16	.37	.08
14	.41	−.11	.34	−.00
15	.38	.03	.31	−.17
16	.34	−.02	.28	.10
17	.32	.13	.26	.11
18	.29	−.00	.24	.01
19	.26	.05	.20	.02
20	.24	−.02	.17	−.06
21	.21	.01	.14	−.01
22	.18	−.11	.11	−.09
23	.16	.11	.09	.10
24	.14	.02	.07	.01
Standard Error[a]	.10	.10	.12	.12

[a] *Individual* standard error under hypothesis that actual autocorrelations are zero at all lags ($T^{-1/2}$ for T observations). The same standard error applies to each autocorrelation coefficient.

Before drawing implications from the model we are interested in estimating δ and testing its statistical significance, in the present context testing the hypothesis that velocity in fact has a tendency to drift in either direction. The least squares results are

$$V_t = V_{t-1} + u_t - .0316 \qquad (3)$$

$$t(\hat{\delta}) = -1.636$$

for the full series and

$$V_t = V_{t-1} + u_t - .0181 \qquad (4)$$

$$t(\hat{\delta}) = -1.094$$

for the shorter series. Using a normal approximation to the t-distribution, the first estimate is just significant at the .102 level against the alternative $\delta = 0$ in a two-tailed test.[6] The estimate for the shorter series

---

merator and denominator of our r_1 we obtain the limit,

$$\lim_{\phi \to 1} E(r_1) = \frac{n-5}{n}$$

Substituting $n = 92$ for our full sample and $n = 70$ for our shorter sample, we obtain .95 and .93, respectively, compared with the corresponding values of .95 and .92 given in Table 1. We also note that r_1 is very nearly the least squares regression coefficient of V_t on V_{t-1} if the first observation is regarded as fixed. To check further on the hypothesis $\phi = 1$, we therefore estimated ϕ using the "backforecasting" technique of Box and Jenkins which is essentially a weighted least squares procedure, taking into account the first observation. These estimates differed from unity only in the third decimal place for both the full and shorter series.

[6] The two-tailed test is the appropriate test in the present context since we have no a priori reason to suppose that the series will drift in either direction.

(which is, of course, not independent of that for the longer series) is just significant at the .274 level in a two-tailed test.

We conclude that by the usual statistical criteria, we cannot reject the hypothesis that the data are generated by a simple random walk model with zero drift:

$$(5) \qquad V_t = V_{t-1} + u_t$$

It is important to emphasize that support for the hypothesis that V_t is a random walk depends only on the lack of autocorrelation in the first differences of the series and *not* on the value of δ. If δ were significantly negative (or positive) we would infer that the series was a random walk with drift. The random walk property per se depends only on the apparent lack of autocorrelation in successive changes in velocity. The nonstationarity of a random walk implies that the series displays no affinity for a mean value, rather it will tend to wander away from any chosen starting point. When δ is zero, it is just as likely to wander in one direction as the other.

To check against the possibility that more complex mechanisms may be required to account for the stochastic structure of the data, it is desirable to test the *overall* smallness of the autocorrelations of velocity changes. This is readily accomplished with use of the Q-statistic due to Box and D. A. Pierce. Denoting the sample autocorrelation at lag j by r_j, the statistic is given by

$$(6) \qquad Q = T \sum_{j=1}^{K} r_j^2$$

where T is the number of observations and K is "sufficiently large." Q is distributed as χ^2 with K degrees of freedom under the null hypothesis that the true autocorrelations are zero. For both series the Q-statistics were smaller than their expected values for $K=12$, 24, and 36, and, thus, the null hypothesis cannot be rejected. As an additional check on the adequacy of the model, we fitted the first-order autoregressive model

$$(7) \quad (V_t - V_{t-1}) = \phi(V_{t-1} - V_{t-2}) + u_t + \delta$$

and the first-order moving average model

$$(8) \qquad (V_t - V_{t-1}) = u_t - \theta u_{t-1} + \delta$$

to both series and found that in both cases the relevant parameters, ϕ and θ, respectively, were small and not significant.[7]

While the absolute changes in velocity are markedly larger prior to 1891, *relative* changes expressed as the changes in the *log* of velocity are not. This suggests that *logs* may provide a variance stabilizing transformation, perhaps eliminating heteroscedasticity present in the full series. The random walk model in *logs* is then

$$(9) \qquad \ln V_t = \ln V_{t-1} + u_t' + \delta'$$

and the estimate of δ' for 1869–1960 and its t-ratio are

$$(10) \qquad \begin{aligned} \hat{\delta}' &= -.0109 \\ t(\hat{\delta}') &= -1.456 \end{aligned}$$

which represents a significance level of about .145 in a two-tailed test against the alternative hypothesis $\delta=0$. Q-statistics for $K=12$, 24, and 36 are all close to their expected values under the null hypothesis that successive *log* changes are uncorrelated.

In summary, our tests based on the Friedman-Schwartz data do not provide evidence against the hypothesis that changes in velocity are serially uncorrelated. There would seem to be only marginal evidence against the further hypothesis that the expected change is zero, particularly for the period following 1891.[8,9]

---

[7] For the estimation theory underlying application of such models, see Box and Jenkins, ch. 7.

[8] Friedman and Schwartz have also tabulated (Table A5, p. 774) a velocity series for the more various definition of the money supply, which includes currency and demand deposits but excludes time deposits. The estimate of δ for this data over the period 1915–60 is $-.004$ with a t-ratio of $-.1379$. The Q-statistics are smaller

III. Postsample Prediction: 1961–70

Further evidence on alternative models for velocity as well as additional insight into their implications are provided by postsample prediction tests for the years 1961–70. It is apparent from equation (2) that the conditional expectation of a future observation V_{t+n}, given the history of the series at time t, denoted by $\hat{V}_t(n)$, is simply

$$(11)\quad \begin{aligned} \hat{V}_t(n) &= E(V_{t+n} \mid \ldots, V_{t-1}, V_t) \\ &= E(V_t + u_{t+1} + \ldots \\ &\quad + u_{t+n} + n\delta \mid \ldots, V_t) \\ &= V_t + n\delta \end{aligned}$$

and this is readily shown to be the minimum squared error prediction. Thus the profile of predictions at time t is simply a straight line with origin V_t at time t and slope δ. Of course, if δ is zero as in (5), then the conditional expectation for any horizon is just V_t. The variance of prediction errors $(V_{t+n} - \hat{V}_t(n))$ is

$$(12)\quad \begin{aligned} &\mathrm{Var}\,(V_{t+n} - \hat{V}_t(n)) = \\ &\quad \mathrm{Var}\,(u_{t+1} + \ldots + u_{t+n}) = n\sigma_u^2 \end{aligned}$$

where σ_u^2 is the variance of the disturbances u. According to the random walk model, then, the only aspect of the history of velocity which is relevant for prediction is the last observation. Uncertainty with respect to the future, as measured by variance of prediction errors, is just proportional to horizon. In the special case $\sigma_u^2 = 0$ we are back to the situation where velocity is constant or at least completely deterministic. In the present case, however, the sample estimates of σ_u^2 are .0340 for the full series and .0189 for the shorter series.

To illustrate prediction of velocity using the random walk models we have computed velocity for the period 1960–70 using a procedure suggested to us by Anna Schwartz. Annual money supply data are averages of monthly data from the *Federal Reserve Bulletin*, pp. 72–77. The income numbers are Department of Commerce National Income multiplied by the factor .882 in order to link this income series with the Kuznets series at 1947.[10] Computa-

---

than their expected values at $K=12$, 24, and 36. Velocity series other than the Friedman-Schwartz data are also of interest. In particular, we have examined the serial correlation structures of two quarterly series over the postwar period. One series is constructed from the standard seasonally adjusted Department of Commerce data as reported in the *Survey of Current Business*, with money defined as currency plus demand deposits and income as gross national product. The second series differs only in that it uses data unadjusted for seasonality. In both cases after appropriate differencing serial correlation is conspicuous only at lags 1, 3, 4, and 5 quarters. Correlation at lag 4 may be accounted for by seasonality (still present in the "seasonally adjusted" series) in the series, at lag 1 by nonseasonal serial correlation, and at lags 3 and 5 by the interaction of seasonal and nonseasonal correlation (see Box and Jenkins, ch. 9). Setting aside seasonal correlation, our preliminary results suggest, then, that only correlation between successive quarters is observed in the data. Regardless of the particular definitions of money and income chosen, the method of aggregation within time periods may be expected to have some effect on the pattern of serial correlation. For example, velocity computed using the money stock at the beginning of each period may differ in this respect from velocity computed using the average money stock over the entire period as it is for all the series we have examined. This is due to a well-known effect of time aggregation discovered by Holbrook Working. We (and others) have proceeded under the implicit assumption that the latter method provides the economically relevant definition of velocity.

[9] Friedman has suggested to us that 1906 might be an appropriate breaking point in the velocity series since after that year the velocity series for the United States and the United Kingdom follow each other more closely than was the case prior to 1906 when *U.S.* velocity was falling rapidly. Although the reference to British data in the choice of a particular breaking point is inconsistent with our objective to focus exclusively on the velocity series, we have calculated $\hat{\delta}$ and t for the subperiods 1896–1906 and 1907–1960. They are $\hat{\delta} = -.061$ and $t(\hat{\delta}) = -1.572$ for 1869–1906 and $\hat{\delta} = -.012$ and $t(\hat{\delta}) = -.633$ for 1907–1960. The t-tests have significance levels .120 and .528, respectively. Similar results were obtained for the *log* of velocity. Surprisingly, this alternative breaking point in the series would not lead us to alter our conclusions with respect to either the random walk hypothesis or the statistical significance of δ; although long-term trends are visually more striking when this breaking point is chosen.

[10] This is clearly only one of a large number of ways to obtain velocity data for 1961 to 1970 that are con-

TABLE 2—VELOCITY OF MONEY 1961–1970: ACTUALS AND PREDICTIONS FROM RANDOM WALK MODELS

Year	Actual	$\hat{V}_{1960}(n)$ $\delta=-.0316$	$\sqrt{n\hat{\sigma}_u^2}$ $\hat{\sigma}_u^2=.0340$	$\hat{V}_{1930}(n)$ $\delta=-.0181$	$\sqrt{n\hat{\sigma}_u^2}$ $\hat{\sigma}_u^2=.0189$	$\hat{V}_t(n)$ $\delta=0$
1961	1.68	1.69	.184	1.70	.137	1.72
1962	1.71	1.66	.260	1.68	.194	1.72
1963	1.69	1.62	.319	1.67	.238	1.72
1964	1.71	1.59	.368	1.65	.274	1.72
1965	1.75	1.56	.412	1.63	.307	1.72
1966	1.75	1.53	.451	1.61	.336	1.72
1967	1.75	1.50	.487	1.59	.367	1.72
1968	1.72	1.47	.520	1.57	.388	1.72
1969	1.73	1.44	.552	1.56	.412	1.72
1970	1.73	1.40	.582	1.54	.435	1.72

tion of predictions for the years 1961–70 from origin date 1960 requires only the number V_{1960} which is 1.72 and alternative estimates of δ, namely $-.0316$ based on the full series, $-.0181$ based on the shorter series, or zero if we wish to accept the null hypothesis $\delta=0$. The alternative predictions along with the realized actuals and estimated standard deviations $n^{\frac{1}{2}}\hat{\sigma}_u$ are presented in Table 2.

The first thing that is apparent from Table 2 is that the movement of velocity during the postsample decade was rather slight by historical standards. From an initial value of 1.72 in 1960, the updated series moved in the range 1.69 to 1.76, ending at 1.73 in 1970. All three series of forecasts are well within one estimated standard deviation at all horizons. This fact reinforces the impression that the series was far less "noisy" during this decade than it had been typically in the past. Further, this excludes the possibility of strong rejection of any version of the random walk model in favor of another purely on grounds of postsample prediction. Nevertheless, we may still make some qualitative comparisons. The version $\delta=-.0316$ results in forecasts which are uniformly below the actuals from 1962 on. The version $\delta=-.0181$ also falls behind the actuals after 1962. The simplest version $\delta=0$ has the smallest absolute errors from 1964 on.

A prediction test which is perhaps more relevant for policy purposes is one-year ahead prediction. The one-step ahead errors and sum of squared errors for the postsample decade are given in Table 3. The smallest sum of squared errors is associated with $\delta=0$, the no-drift model. The fact that all of the sums of squared errors are very much less than either estimate of σ_u^2 adds to the impression that the velocity series was unusually quiet during the 1960's. (Recall that the one-step ahead error $[V_{\tau+1}-\hat{V}_\tau(1)]$ is simply $u_{\tau+1}$.) Tests of the hypothesis of equality of variance of velocity changes for 1961–70 versus 1869–1960 and 1891–1960 both lead to rejection at a very high level (less than .01) of significance.

The qualitative conclusions of this section are unaltered if the *log* of velocity is compared with predictions of model (9)–(10) and a zero drift model. The dispersion of relative changes in velocity also appears to have diminished in the 1960's, equality of variance for 1960–70 versus 1869–1960

---

sistent with the Friedman-Schwartz series. We have tried many other methods and everyone of them has led to similar conclusions in the postsample prediction tests.

TABLE 3—VELOCITY OF MONEY 1961–1970: ACTUALS AND ONE-YEAR-AHEAD PREDICTIONS FROM RANDOM WALK MODEL

Year of Actual	Actual	$\hat{V}_t(1)$ $\delta=-.0316$	$V_{t+1}-\hat{V}_t(1)$	$\hat{V}_t(1)$ $\delta=-.0181$	$V_{t+1}-\hat{V}_t(1)$	$\hat{V}_t(1)$ $\delta=0$	$V_{t+1}-\hat{V}_t(1)$
1961	1.68	1.69	−.01	1.70	−.02	1.72	−.04
1962	1.71	1.65	.06	1.66	.05	1.68	.03
1963	1.69	1.68	.01	1.69	.00	1.71	−.02
1964	1.71	1.66	.05	1.67	.04	1.69	.02
1965	1.75	1.68	.07	1.69	.06	1.71	.04
1966	1.76	1.72	.04	1.73	.03	1.75	.01
1967	1.75	1.73	.02	1.74	.01	1.76	−.01
1968	1.72	1.72	.00	1.73	−.01	1.75	−.03
1969	1.73	1.69	.04	1.70	.03	1.72	.01
1970	1.73	1.70	.03	1.71	.02	1.73	.00
Sum of Squared Errors			.0157		.0105		.0061

being rejected at a very high level (less than .01) of significance.

In summary, one would have done very well during 1961–70 to have predicted income by simply using the velocity figure just observed for the past year. This is in contrast to the procedure that would have been recommended by the assumption underlying regression models of the money-income relationship. In particular, these models are of the form

$$Y_t = VM_t + \epsilon_t \tag{13}$$

If the standard deviation of ϵ_t is proportional to M_t then

$$\frac{Y_t}{M_t} = V + (\epsilon_t/M_t) \tag{14}$$

is a properly specified regression model and the least squares estimate of V is simply

$$\hat{V} = \bar{V} = \frac{1}{T}\sum_{t=1}^{T}\frac{Y_t}{M_t} \tag{15}$$

the mean velocity over the sample period. For the Friedman-Schwartz data we have

$$\hat{V} = 2.42, \quad 1869\text{–}1960 \tag{16}$$

and

$$\hat{V} = 1.94, \quad 1891\text{–}1960 \tag{17}$$

both of which would have provided very much poorer predictions than any of the random walk models for the period 1961–70.

IV. Implications of the Random Walk Model

We are obliged to conclude from the foregoing analysis that the Friedman-Schwartz velocity series, when considered as a univariate stochastic process, is empirically indistinguishable from a random walk; that is, the simple summation of uncorrelated shocks. The fact that it appears *ex post* that nonrandom patterns and trends occur in this series is not inconsistent with the hypothesis that the series is a random walk. As William Feller notes in his discussion of apparent trends in random walks,

> Most stochastic processes in physics, economics and education are of this nature, and our findings should serve as a warning to those who are prone to discern secular trends and deviations from average norms. [p. 68][11]

---

[11] It was once (and to some extent still is) quite popular to fit polynomials in time to economic data series and the temptation to do so is understandable, given the strong *ex post* patterns that can be seen in the velocity series. As part of the analysis, we fitted a quadratic in time to the shorter velocity series using ordinary least squares. Taken at face value the t-ratios of the coefficients of time and time squared in this regression are

This warning and the hypothesis that velocity is a random walk does not mean that there is no meaningful relationship between the money supply and national income. It does imply, however, that one must exercise particular caution in drawing inferences about this relationship solely on the basis of the historical behavior of velocity.[12] In particular, we have seen that there is no strong statistical evidence from the series itself for attributing the decline in velocity observed over the period 1869–1946 to an underlying tendency to drift in a negative direction. More important for short-term forecasting and policy evaluation, it would appear fruitless to attempt to identify apparent "deviations" from "trend" since, if velocity is a random walk, there is no reason to expect that any such deviations will be corrected in the future. A random walk process exhibits no affinity for a mean level or particular trend line, but rather will display ever larger departures from any given level or trend line as time progresses.

It is a matter of some interest to consider whether our findings can be reconciled with some of the leading hypotheses regarding the fundamental nature of the money-income relationship. For example:

H_0: Velocity fluctuates around a fixed mean (the classical case). This would be implied by an *LM-IS* system under classical assumptions of flexible wages and prices if disturbances are monetary shocks which shift the *LM* curve. In equation form the hypothesis states

$$V_t = V + \tilde{V}_t \tag{18}$$

where $\{\tilde{V}_t\}$ is a mean-reverting stationary error and V is a constant.

It is clear that the random walk model is inconsistent with H_0 due to the nonstationarity of V_t.

H_1: Velocity is constant if properly defined as the ratio of "true" income to true money. The discrepancies between measured and true magnitudes may be due to measurement errors or transitory components in money and income. For true magnitudes $\bar{Y}_t$ and $\bar{M}_t$, we have then

$$\frac{\bar{Y}_t}{\bar{M}_t} = V \tag{19}$$

This hypothesis is essentially the same as H_0 since both refer to transitory departures from a fixed long-run relationship. Assuming that measured quantities are related to true quantities by

---

highly significant, but the residuals are strongly autocorrelated. This is not evidence against the random walk hypothesis, however, since it is typically the case that any particular section of the history of a random walk can be well approximated by a polynomial in time. Using Monte Carlo techniques, we generated numerous random walks with error terms drawn from a normal distribution having parameters of our estimated model. We consistently found that time and time squared had "statistically significant" coefficients in least squares polynomial regressions of these known random walks. This is an example of model misspecification in which the irrelevant included variables (time and time squared) are correlated with the relevant excluded variables (in this case lagged velocity). When lagged velocity is added to the regression for the actual velocity series, all the coefficients have large t-ratios and the autocorrelation of the residuals is reduced substantially. This result appears at first to be strong evidence contrary to the random walk hypothesis. Such a conclusion is not warranted. In this situation ordinary least squares results are misleading because the classical assumptions are not satisfied. Once again we found that we were repeatedly able to replicate these results for random walks generated in Monte Carlo experiments. The finding that time and time squared show up significantly in these Monte Carlo experiments is an interesting result with potentially important implications for the analysis of economic time-series. We plan to examine this question in a later paper. It also should be added that post-sample predictions with the time-polynomial model are far worse than the random walk model.

[12] Milton Friedman has pointed out to us that over the past 90 years velocity in the United States and the United Kingdom is correlated on a year-by-year basis and that the correlation coefficient is higher than could be attributed to chance. This is also true if year-to-year changes in the two velocity series are correlated. Such a finding does not vitiate the conclusion that the *U.S.* (and possible the *U.K.*) series is a random walk. It simply suggests that both series may be subject to random shocks which are cross-correlated.

$$(20)\qquad \begin{aligned} Y_t &= \bar{Y}_t \cdot e^{\tilde{Y}_t} \\ M_t &= \bar{M}_t \cdot e^{\tilde{M}_t} \end{aligned}$$

where $\tilde{Y}_t$ and $\tilde{M}_t$ are zero mean random errors. Then if $V_t = Y_t/M_t$ we have

$$(21)\qquad \ln V_t = \ln V + (\tilde{Y}_t - \tilde{M}_t)$$

which is equivalent to (18) except it is in *log* form.

H_2: The demand for real per capita money balances is a function of permanent real per capita income, see Friedman (1959).

Friedman shows on p. 336, equation (7), that this hypothesis implies that *measured* velocity, the ratio of measured income to measured money, is given by

$$(22)\qquad \begin{aligned} V_t = \gamma &+ (Y_t - Y_t^p) \\ &+ (1 - \eta)(Y_t^p - N_t) \end{aligned}$$

where all variables are expressed in *logs*, all monetary variables are deflated by permanent price level, superscript p denotes permanent magnitudes, N_t denotes the *log* of population, γ is a constant, and η (Friedman's δ) is the elasticity of the demand for real permanent balances with respect to real permanent per capita income. Friedman estimated η to be 1.81.

We begin our discussion of (22) by noting that the last term, equivalent to permanent velocity in Friedman's framework apart from a constant, may reasonably be argued to follow a random walk if Y^p is interpreted as long-run expected income. Under that interpretation it follows that changes in Y^p should be essentially uncorrelated, or else economic agents would be able to anticipate future revisions in their expectations merely by examining past revisions. In other words, changes in Y^p are reasonably assumed to be unanticipated.[13] The nature of N_t is not crucial since we may either take it to be a deterministic function of time or simply rephrase the above argument in terms of per capita permanent income. Clearly then, if our data consisted of observations on *permanent* velocity, the model would lead us to expect random walk behavior. It would also predict a negative mean for the first differences since $(1-\eta)$ is negative $(-.81)$ and real per capita permanent income presumably grows over time. The stochastic properties of measured velocity, as we shall see below, depend crucially on the magnitude of η and the stochastic structure of the measured income series.

Friedman assumed, at least for empirical purposes, that an appropriate estimate of permanent income is expected income given by[14]

$$(23)\qquad Y_t^p = \beta \sum_{i=0}^{\infty} (1 - \beta)^i Y_{t-i} + \alpha, \qquad 0 \leq \beta \leq 1$$

where α allows for a long-term drift in income to be reflected in expectations. As John Muth has shown, (23) will be optimal (minimum mean square error) if Y_t is generated according to

$$(24)\qquad Y_t = Y_{t-1} + \epsilon_t - (1 - \beta)\epsilon_{t-1} + \alpha\beta$$

where $\{\epsilon_t\}$ is a sequence of zero-mean uncorrelated disturbances.[15] Note that according to (24) the first differences of Y_t will in general be autocorrelated at lag one only. Using (24) it is readily demonstrated that the change in expected income is given by

---

[13] We offer this only as one plausible interpretation of the model in the interest of seeing if it is consistent with observed behavior of measured velocity. Other interpretations are clearly possible.

[14] Equation (23) is a discrete version of the continuous weighting scheme given in Friedman (1959, p. 337). We use (23) as part of the empirical framework for our subsequent analysis but with great reluctance because it is far from clear that such a method would be used by rational economic agents to form expectations.

[15] To our knowledge, there is no evidence that the process generating income is of the form of (24), or equivalently, that (23) is an appropriate extrapolative forecasting mechanism.

$$(25)\qquad (Y^p_t - Y^p_{t-1}) = \beta\epsilon_t + \alpha\beta$$

which is a random process by the definition of $\{\epsilon_t\}$ with mean $\alpha\beta$. This establishes that changes in permanent velocity will also be uncorrelated.

Noting that permanent velocity differs from measured velocity by the term $(Y_t - Y^p_t)$, which according to (24) is just $(1-\beta)\epsilon_t - \alpha\beta$, we now may write an expression for the first differences of measured velocity in terms of the random disturbances $\{\epsilon_t\}$ and parameters of the model, namely from (22)

$$\begin{aligned}(26)\quad (V_t - V_{t-1}) = \\ &(1-\beta)(\epsilon_t - \epsilon_{t-1}) \\ &+ (1-\eta)\beta\epsilon_t + (1-\eta)(\alpha\beta - r)\end{aligned}$$

where r is the rate of growth of population (assumed constant) and hence $(\alpha\beta - r)$ the rate of growth of per capita Y^p. Rearranging we have

$$\begin{aligned}(27)\quad \Delta V_t = \{1 - \beta\eta\}\epsilon_t - (1-\beta)\epsilon_{t-1} \\ + (1-\eta)(\alpha\beta - r)\end{aligned}$$

The differences of V_t will clearly be autocorrelated in general although the magnitude of the autocorrelation will depend on values of η and β. In particular, the first-order autocorrelation coefficient is

$$(28)\qquad \rho = \frac{(1-\beta\eta)(\beta - 1)}{(1-\beta\eta)^2 + (1-\beta)^2}$$

and all higher order autocorrelation coefficients are zero. It is easily seen from (28) that when

$$(29)\qquad \beta\eta = 1$$

then ρ is zero. Keeping in mind that Friedman estimated η to be 1.81 and β to be .4 it is interesting to note that pairs of β and η consistent with $\rho = 0$ include

η:	2.50	2.25	2.00	1.81	1.75	1.5
β:	.4	.44	.50	.55	.57	.66

The value of ρ will also be small for values of β and η in the neighborhood of Friedman's point estimates even though they do not satisfy (29) exactly. For example, fixing η at 1.81 we have

β:	.4	.5	.55	.6
ρ:	$-.33$	$-.19$	.0	$+.20$

and alternatively fixing β at .4 we have

η:	1.81	2.00	2.25	2.50
ρ:	$-.38$	$-.30$	$-.16$	.0

The point of these examples is simply that there are plausible values of β and η which are well within presumed sampling error that imply we should find very little autocorrelation of any order in velocity.

It is also interesting to note that the expected change in the *log* of measured velocity in (22) will be $(1-\eta)$ times the expected rate of change in permanent real per capita income. Taking as an estimate of the latter, the historical mean rate of growth of real per capita income of about .02 we have a derived expected rate of change of velocity of $-.0162$ using Friedman's value of 1.81 for η. We note that this derived estimate is within one standard error of our estimate of δ' although it is more than two standard errors (on $\hat{\delta}$) from zero.

In short, we are able to find an interpretation of Friedman's 1959 model that would be consistent with the finding that measured velocity is a random walk.

***H*$_3$: Money and income are both nonstationary upward-drifting, but essentially unrelated, time-series (the strict Keynesian case).

Under this hypothesis velocity is simply a statistical artifact of the histories of money and income, the *log* of velocity evolving according to

$$(30)\qquad ln\ V_t = ln\ Y_t - ln\ M_t$$

Since $ln\ Y_t$ and $ln\ M_t$ are both nonstationary, it is readily argued that their difference will be nonstationary under this hypothesis, for example, if both are nonconvergent summations of random errors. Their difference, however, need not be a random walk. Whether there is a drift in $\ln V_t$ would depend on the relative rates of drift of $\ln Y_t$ and $\ln M_t$. The strict Keynesian hypothesis may therefore also be consistent with our empirical findings.

H_4: Income is a distributed lag on money and possibly other variables with fixed coefficients (see, for example, Leonall Andersen and Keith Carlson and Arthur Laffer and R. David Ranson).

The simplest example of this class of models would be of the form

$$Y_t = \sum_{i=0}^{k} \beta_i M_{t-i} \tag{31}$$

with the β_i being fixed coefficients. The model may be rewritten in the form

$$\frac{Y_t}{M_t} = \beta_0 + \sum_{i=1}^{k} \beta_i \frac{M_{t-i}}{M_t} \tag{32}$$

A random walk in V_t would imply that the weighted sum of ratio changes in money be a random walk. For instance, if $k=1$ then M_{t-1}/M_t, roughly one plus the percentage change in M, must evolve according to a random walk. The other possibility is that the coefficients β_i are not stable. The conditions imposed on the model are, of course, much more complex if additional explanatory variables, such as government purchases of goods and services, are included, as is often the case.

H_5: Velocity is determined in a general *LM-IS* framework where unemployment equilibria are possible and/or the *IS* curve shifts.

It should be clear that in the context of the *LM-IS* framework, any of the circumstances allowing an unemployment equilibrium or any shift in the *IS* curve, including those due to expectations of the rate of inflation, could be compatible with nonstationary behavior in velocity and therefore not inconsistent with our random walk model.[16]

The general conclusion which emerges from review of the six hypotheses reviewed above is that, while some are clearly inconsistent with a random walk in velocity, there remain both monetarist and Keynesian alternatives which may be reconciled with the hypothesis that velocity is a random walk.

---

[16] William Poole has suggested another interesting hypothesis, namely, that velocity may be a function of the "permanent" interest rate and the permanent interest rate may be measured by the rate on perpetual bonds. The price of a perpetual bond is $1/i$ and given the evidence that security prices follow a random walk we might expect velocity to do so also.

REFERENCES

L. C. Andersen and K. M. Carlson, "A Monetarist Model for Economic Stabilization," *Fed. Reserve Bank St. Louis*, Apr. 1970, *52*, 7–25.

G. E. P. Box and G. M. Jenkins, *Time Series Analysis*, San Francisco 1970.

——— and D. A. Pierce, "Distribution of Residual Autocorrelations in Autoregressive Integrated Moving Average Time Series Models," *J. Amer. Statist. Assn.*, Dec. 1970, *65*, 1509–26.

W. Feller, *An Introduction to Probability Theory and Its Applications*, New York 1957.

M. Friedman, "The Quantity Theory, A Restatement," in his *Studies in the Quantity Theory of Money*, Chicago 1956.

———, "The Demand for Money: Some Theoretical and Empirical Results," *J. Polit. Econ.*, Aug. 1959, *66*, 327–51.

———, *Dollars and Deficits*, Englewood Cliffs 1968.

——— and A. J. Schwartz, *A Monetary History of the United States 1867–1960*, Princeton 1963.

M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, vol. 3, New York 1966.

A. B. Laffer and R. D. Ranson, "A Formal Model of the Economy," *J. Bus., Univ. Chicago,* July 1971, *44,* 247–70.

J. F. Muth, "Optimal Properties of Exponentially Weighted Forecasts," *J. Amer. Statist. Assn.,* June 1960, *55,* 299–306.

H. Working, "A Note on the Correlation of First Differences of Averages in a Random Chain," *Econometrica,* Oct. 1960, *28,* 916–13.

Board of Governors of the Federal Reserve System, *Fed. Res. Bull.,* vol. 59, Washington 1973.

U.S Office of Business Economics, *Surv. Curr. Bus.,* Washington, various issues.

Faculty Salaries: Is There Discrimination by Sex, Race, and Discipline?

By NANCY M. GORDON, THOMAS E. MORTON, AND INA C. BRADEN*

Many variables which affect salary, such as age, education, and race, have been studied in detail. Less is known about salary profiles and differentials related to sex. This study uses data on all full-time academic employees of a large urban university. It controls for age, race, years at the university, education, rank, and department. We investigate the relationship between salary and the explanatory variables for men and women separately as well as for all employees together, where sex is an additional independent variable. Regression analyses yield unambiguous results; about three-fourths of the variation in salary across employees is explained. The salary profile for women differs from that for men primarily in the effects of race, age, and department. Salary increments between ranks are similar. However, on average, a woman faculty member earns 11 percent less than would be predicted for a man with her characteristics. Blacks earn 13 percent more than comparable whites. A medical school faculty member with an M.D. earns 93 percent more than an otherwise comparable Ph.D. in the social sciences.

I. Overview

When considering theories of a firm's salary structure one can proceed at three levels: casual observation, formal modelling, and empirical investigation. All three levels are necessary and ideally interact to further the development of any area. Many variables affecting salary have been studied, for example, age, education, and race. Less is known about the effect of the employees' sex. To be sure, there has been a great deal of "armchair empiricism," and a fair amount of formal modelling, some originally formulated with respect to race but applicable to sex discrimination as well. However, there is little work on wage profiles by sex or differentials between men and women which controls for the a priori relevant variables.

The most satisfactory way to control for such factors is to build a wage structure model encompassing them and to fit it empirically. While one is never able to measure all the relevant variables, the more that can be included,[1] the more dependable the results. For this study, we have detailed information on virtually every full-time academic employee of a large urban university. In particular, we estimate salary as a function of age, race, years at the university, current education, rank, and department for men and women separately, as well as for all employees together, where sex is an additional independent variable.

The coefficients of the independent variables in our regressions have high t-values and the R^2s are approximately 0.78. The salary profile for women differs from that for men primarily in the effects of race, age, and department. Salary increments between ranks are similar. However, the average woman faculty member earns 11.4 percent less than would be predicted by the male regression for a man with her characteristics. If one is willing to assume that the academic labor market in which this university competes is efficient, then one would conclude that women's accomplishments are less than those of otherwise comparable men. On the other hand, if one believes that women face barriers to entry that men do not—that is, if to attain the same position a woman must be better than a man—then one would conclude that our estimates of the differentials are biased downwards. Unfortunately, the data that would allow us to distinguish between such hypotheses are unavailable.

---

* Assistant professor of economics, associate professor of industrial administration, Carnegie-Mellon University, associate professor of psychology, Allegheny County Community College, respectively. We wish to thank David Cass, Robert Lucas, Timothy McGuire, and Leonard Rapping for their valuable suggestions and Margaret Hart and Marcel Weiner for their diligence as research assistants. Roman Weil, Lester Lave, and our programmer Mark Gelfand were especially helpful.

[1] Assuming that the correct model, including interactions, has been specified.

The following section reviews past research in this area. Section III discusses our data, preliminary analysis, and model. Section IV presents our empirical results. Section V is a summary.

II. Literature Survey

Although attitudes have come a long way since F. Y. Edgeworth worried about whether women *should* receive equal pay for equal work, little empirical evidence has been accumulated on whether or not they do. Factors affecting the labor force participation of women have been investigated to some extent.[2] The Women's Bureau of the Department of Labor (*WBDL*) has prepared much data of the cross-tabulation form.[3] Henry Sanborn adjusted data of a similar nature and estimated a differential of 13 percent. Valarie Oppenheimer observed that most working women hold jobs predominantly filled by women and that these jobs pay less than "male" jobs requiring similar education. Francine Weisskoff and Harriet Zellner also addressed the question of occupational segregation.

Mary Hamilton analyzed survey data from the Chicago area for four occupations and found that after controlling for age, education, experience, training, seniority, and firm differences, the estimated wage differentials due to sex varied between 8 and 18 percent. Malcolm Cohen, also using survey data, concluded that the major reason women were paid less than men was that they were concentrated in the lower paying occupations! Using large scale survey data, Ronald Oaxaca controlled for many factors but *excluded* job type. As expected, he found much larger differentials—54 percent between white males and white females.

The only theoretical models on discrimination by sex known to us are by Barbara Bergmann, Gordon and Morton, and Janice Madden. The first discusses the crowding (and subsequent lowering of marginal productivity) of minority group members into those few occupations which possess no barriers to entry. The second emphasizes that a low mobility group may be paid less for comparable work due to a combination of monopsonistic competition and fellow worker distaste. The third investigates similar assumptions in both competitive and noncompetitive markets.

III. Data, Preliminary Analysis, and Model

Detailed information for each of the academic employees of the university as of October 1970 was obtained from the personnel department. The precise number of observations is suppressed to preserve the anonymity of the university, but the sample size is between 1,000 and 2,000, of whom 18 percent are women. Elimination of incomplete or inconsistent observations reduced the total sample by eleven men and two women.

The information for each faculty member included: date of birth; date employed by the university; sex; race; educational degrees held; academic rank; academic department; academic salary.

A description of the modified variables used in the study follows. *Age*, as of October 1970 is computed to the nearest 0.1 year. *Seniority*, defined as the amount of time employed at the university by October 1970, is also computed to the nearest 0.1 year. *Sex* is either male or female. *Race* was aggregated to white, black, and other. *Educational degree* was aggregated to bachelor's degree, master's degree, doctoral degree, doctor of dentistry, and law degree. Information on field of degree and multiple degrees was available but judged difficult to use. *Academic rank* was aggregated to instructor, lecturer, teacher, assistant professor, associate professor, professor, clinical assistant professor, and clinical associate professor. The data included more detailed academic ranks (for example, visiting associate research professor) but they did not differ significantly from the corresponding group (for example, associate professor). However, although the number of faculty with clinical ranks is small, they did have distinctly higher salaries and they are not expected to conduct research. Thus they were kept separate. *Academic department* was aggregated to fine arts, engineering and science, social science, quantitative social science, education, medicine, health, law, and other. These are the university's classifications, except for two cases. Based on preliminary analyses which showed them to be similar, the business school, the economics department, and the school of public affairs were grouped and designated as quantitative social science; for the same reason, the dental, health, and nursing departments were grouped and designated by health. *Academic salary* represents the usual 9-month payment except in a few cases where the basis was 10 or 11 months. Salary is measured in hundreds of dollars.

---

[2] See Eli Ginzberg, Marylyn Goldberg, J. N. Hedges, Gertrude McNally, Charlotte Phelps, E. Waldman, and Harold Wilensky.

[3] See references from the *WBDL*.

Since the "true" relationship between salary and the independent variables described above is complicated and since the theories reviewed in the previous section did not suggest a particular specification, extensive one-, two-, three- and four-way cross-tabulations were undertaken to isolate the important non-linear effects. A piecewise linear function for age proved necessary. There seemed to be significant interactions between sex and such variables as seniority, race, and department. For this reason, separate models were fit for each sex, as well as a model for all employees in which sex was an additional (dummy) variable. Other interactions (for example, between education and department) were judged unimportant and were suppressed.

We turn now to the formal model. Both salary and the logarithm of salary were employed as dependent variables. The former would fit salary as a linear function of the independent variables, the latter as a multiplicative function. The two specifications performed equally well from the point of view of the percentage of explained variation.[4] Since percentage deviations[5] are more interesting than dollar differences, only the logarithmic specification will be discussed.

Model I fits a common salary structure for both sexes (by ordinary least squares regression):

$$\text{(1)} \quad \ln W_i = a_0 + f(A_i) + a_5 L_i + d^n(E_i) + d^m(J_i) + d^q(S_i) + d^r(R_i) + d^t(D_i) + e_i$$

where, for the ith employee, W_i represents salary, A_i is age, L_i is time at the university (seniority), E_i is education, J_i is academic rank, S_i is sex, R_i is race, D_i is department, and e_i is the disturbance term. All variables are discrete (dummy variables) except age and seniority. The superscripts n, m, q, r, and t stand for, respectively, the number of educational levels, the number of ranks, the number of sexes, the number of race categories, and the number of departments.

The piecewise linear, continuous age function found to be appropriate is:

$$\text{(2)} \quad f(A_i) = a_1 A_i + a_2(A_i - 35)^+ + a_3(A_i - 45)^+ + a_4(A_i - 55)^+$$

(The superscript $+$ means the maximum of the quantity in parentheses and zero.) The function d represents the dummy variables used to include the discrete variables in the equation. For example

$$\text{(3)} \qquad d^n(E_i) = \sum_{j=1}^{n-1} a_6^j E_i^j$$

where n is the number of possible education levels, a_6^j is the coefficient associated with the jth education level and E_i^j is a dummy variable. In particular E_i^j equals one if the ith employee's highest attained educational level is j and equals zero otherwise. Thus for

---

[4] For example, the linear model using male employees results in an R^2 of 0.783 and the logarithmic model in an R^2 of 0.790. When we calculate the percentage explained variation around the curved function in the original linear space, we obtain an R^2 of 0.785.

[5] In the multiplicative model, note that the correct percentage difference from the excluded variable for an estimated coefficient of x is $e^x - 1$. However, for small absolute values of x, x approximates $e^x - 1$ well. The authors present the correct values.

TABLE 1—ESTIMATED COEFFICIENTS FOR MODELS I AND II[a,b]

	All Employees			Male Employees			Female Employees		
	Mean	Coefficient	*T*-Value	Mean	Coefficient	*T*-Value	Mean	Coefficient	*T*-Value
Intercept	—	4.846	258.9	—	4.848	230.3	—	4.816	121.4
Age (Per Year)									
≤35	0.354	0.0060	3.8	0.348	0.0055	3.1	0.381	0.0092	2.7
35+—45	0.315	0.0050	2.8	0.319	0.0058	2.8	0.300	0.0003	0.1
45+—55	0.207	−0.0024	−1.2	0.208	0.0010	0.4	0.201	−0.0137	−3.3
55+	0.124	−0.0014	−1.7	0.125	−0.0020	−2.0	0.118	−0.0002	−0.1
Seniority (Per Year)	6.00	−0.0049	−5.4	6.35	−0.0056	−5.7	4.36	−0.0032	−1.5
Race									
White	0.916	Standard	Standard	0.916	Standard	Standard	0.913	Standard	Standard
Black	0.043	0.127	5.5	0.039	0.139	5.1	0.065	0.075	1.9
Other	0.041	0.005	0.2	0.045	−0.002	−0.1	0.022	0.039	0.6
Education									
Bachelor's Degree	0.043	−0.093	−3.5	0.032	−0.057	−1.7	0.096	−0.186	−4.3
Master's Degree	0.230	−0.068	−4.4	0.177	−0.078	−4.2	0.486	0.083	−2.9
Doctoral Degree	0.541	Standard	Standard	0.581	Standard	Standard	0.338	Standard	Standard
M.D.	0.146	0.361	16.1	0.163	0.372	15.1	0.068	0.264	5.0
D.D.M.	0.023	0.028	0.8	0.026	0.038	1.0	0.012	0.011	0.1
J.D.—L.L.B.	0.017	−0.043	−0.8	0.021	−0.053	−0.9	None	—	—
Rank									
Instructor	0.167	−0.190	−11.4	0.127	−0.201	−10.0	0.356	−0.184	−6.6
Assistant Professor	0.326	Standard	Standard	0.317	Standard	Standard	0.363	Standard	Standard
Associate Professor	0.212	0.185	13.0	0.227	0.190	12.0	0.146	0.192	5.8
Professor	0.242	0.495	29.1	0.278	0.494	26.3	0.068	0.421	9.0
Lecturer	0.034	−0.053	−1.8	0.031	−0.024	−0.7	0.046	−0.098	−2.0
Teacher	0.005	−0.232	−3.5	0.005	−0.220	−2.7	0.009	−0.230	−2.2
Clinical Assistant Professor	0.008	0.284	5.5	0.008	0.291	4.9	0.009	0.303	2.9
Clinical Associate Professor	0.006	0.378	6.3	0.007	0.433	6.7	0.003	−0.038	−0.2
Department									
Fine Arts	0.110	0.002	0.1	0.112	−0.005	−0.2	0.102	0.002	0.0
Engineering and Science	0.191	0.139	7.3	0.219	0.129	6.1	0.059	0.175	3.6
Social Science	0.073	0.096	4.3	0.078	0.082	3.2	0.053	0.191	3.8
Quant. Social Science	0.056	0.186	7.5	0.064	0.176	6.6	0.019	0.142	1.9
Education	0.114	0.189	9.0	0.108	0.182	7.5	0.139	0.160	4.1
Medicine	0.186	0.402	16.2	0.202	0.392	14.0	0.108	0.381	7.5
Health	0.157	0.188	9.7	0.104	0.162	6.6	0.406	0.212	6.6
Law	0.011	0.159	2.2	0.013	0.158	2.1	None	—	—
Other	0.012	Standard	Standard	0.100	Standard	Standard	0.114	Standard	Standard
Sex									
Male	0.825	Standard	Standard	—	—	—	—	—	—
Female	0.175	−0.095	−7.0	—	—	—	—	—	—
Mean *Log* Salary		5.145			5.202			4.881	
Standard Deviation		0.426			0.425			0.318	
Observations		$N(1000<N<2000)$			$0.825N$			$0.175N$	
Predicted Salary "Standard" Employee		\$12,980			\$12,990			\$12,510	
Standard Error of Estimate		0.193			0.196			0.163	
R^2 (Adjusted)		0.794			0.786			0.735	

[a] The regression coefficients have approximate interpretations as the percentage salary differential (divided by 100) from the standard employee, and the standard deviation of the dependent variable and the standard error of the estimate as percentage differentials (divided by 100) from the mean of *log* salary. The predicted dollar salary of the standard employee was calculated as explained in fn. 13.

[b] The independent variables are listed in the first column. The mean values, coefficients, and *t*-values (not standard errors) for the all-employee regression are given in the next three columns, followed by the same three columns for the male-employee regression and then by the same three columns for the female-employee regression. The means of the dummy variables give directly the fraction of the sample having the corresponding characteristic. The proportions of the sample in the four age categories have replaced the means of the age variables which were felt to be less interesting. Note that the mean of the seniority variable gives the average length of employment at the university. The standard employee is a 35-year old, newly hired, white assistant professor in the "other" department with a Ph.D. The standard person is male in the all-employee regression.

each employee only one term representing education actually enters the equation.

Model II, which fit separate specifications by sex, is illustrated below:

$$(4)\quad \ln W_{si} = a_{s0} + f_s(A_{si}) + a_{s5}L_{si} + \overset{n}{d}_s(E_{si}) + \overset{m}{d}_s(J_{si}) + \overset{r}{d}_s(R_{si}) + \overset{t}{d}_s(D_{si}) + e_{si}$$

That is, there are two submodels, one for $s=0$ (males) and one for $s=1$ (females).

The results of Models I and II are presented in Table 1 and are discussed in the following section. Results of other investigations are also included.[6]

---

[6] The corresponding tables are available from the authors.

IV. Empirical Results

To summarize briefly, we find that women's salaries peak earlier, on average, than do men's. For employees of the same age, longer length of employment with the university is associated with lower salaries. Black faculty members earn more than otherwise comparable whites. Except for men with bachelor's degrees (who earn more than those with master's), salary increases with educational degrees and with academic rank. The fine arts department has significantly lower, and the medical school significantly higher, salaries than the rest of the university. Women earn approximately 10 percent less than comparable men. In the following sections, we discuss the separate salary profiles for men and women and then compare them.

A. *Demographic Variables: Age, Seniority, and Race*

The age variable in cross-sectional studies is often used to represent a combination of physical age and seniority, and is in addition confounded whenever skills listed as the same are different. For example, an educational degree awarded twenty years ago is not the equivalent of a recent one. However, since we know how long ago each individual was hired, some of these effects can be removed by using both variables. Thus age is more likely to reflect the fact that income peaks prior to retirement. We used a four-section piecewise linear relationship between salary and age to approximate the non-linear function. The segments (of less than or exactly 35, older than 35 but no older than 45, older than 45 but no older than 55, and older than 55 years of age) were chosen to ensure a reasonable number of observations in each one. We found that men's salaries increase at approximately 0.6 percent per year until they are 45, then at 0.1 percent per year until they are 55 and decrease[7] at approximately 0.2 percent per year over 55 ($t=3.1$, 2.8, 0.4, -2.0, respectively). For women, salaries increase at approximately 0.9 percent per year until they are 35, remain the same until they are 45, decrease at approximately 1.4 percent per year between the ages of 45 and 55, and remain the same thereafter ($t=2.7$, 0.1, -3.3, -0.1, respectively). That is, on average, men's salaries peak approximately fifteen years later than women's salaries.

The effect of each additional year's employment at the university was negative for both men (-0.6 percent) and women (-0.3 percent) ($t=-5.7$, -1.5, respectively); the coefficients were not significantly different from each other. Three observations may explain this result. First, in times of increasing starting salaries, new entrants often earn as much as those with two or three years' experience. The "wage-compression" phenomenon is well known.[8] Second, skills are less likely to be obsolete if an employee has been hired, and hence evaluated, recently. Third, there may be positive returns to job search. Note that for a continuing employee age and seniority progress concurrently.

Both black men and black women[9] earn significantly more than comparable whites, 14 percent for men and 7.5 percent for women ($t=5.1$, 1.9, respectively). A reasonable explanation is the interaction between the recent sharp increase in demand for black faculty and the small available supply.[10]

B. *Academic Variables: Education, Rank, and Department*

Cross-tabulations indicated that the acquisition of a degree has a noticeable impact on the employee's salary, whereas partial fulfillment of the requirements does not.

---

[7] Note that an individual employee's salary is unlikely to decrease over time. However, in an inflationary period salary increments at less than the average rate would result in a lower relative salary, which would appear in cross-sectional data as an absolute decline.

[8] See M. Sami Kassem.

[9] Sample sizes for Spanish surname, Oriental and American Indian faculty are too small to provide any information.

[10] Data collected earlier included marital status but not race. The income profiles derived from this data indicate that single men earn approximately 3 percent less than comparable married men whereas single women earn approximately the same as comparable married women.

Using dummy variables to represent the highest degree earned, we found that those employees with master's degrees earn 2.5 percent more than those with bachelor's degrees; Ph.D. holders earn 7.4 percent more than those with master's degrees; and M.D.s earn 35.5 percent more than those with Ph.D.s.[11] Male faculty with Ph.D.s earn 7.8 percent more than those with master's degrees and M.D.s 37.2 percent more than Ph.D.s. Women faculty with master's degrees earn 10.3 percent more than those with bachelor's degrees, Ph.D.s earn 8.3 percent more than those with master's degrees, and M.D.s earn 26.4 percent more than Ph.D.s. An anomalous result is that men with bachelor's degrees earn only 5.7 percent less than Ph.D.s; that is, they earn 2.1 percent more than those with master's degrees. We observe that there are several male full professors in this category.[12]

The male and female percentage salary increments between ranks are similar, except that men earn a larger increment from associate to full professor. The increments are derived from Table 1 and shown in Table 2. The difference in the increment between associate professor and professor may result from different distributions of rank for men and women. That is, the average time lapse since promotion to full professor may be significantly less for women than for men.

Both men and women in the fine arts department are the least well paid (approximately the same as the standard department, other) and those in the medical school are the highest paid, relative to members of the same sex in different departments. Men in social science do poorly relative to other men, whereas for women it is one of the best paid departments, relative to other women. On the other hand, male faculty members in education are the second highest paid relative to their male colleagues, whereas women faculty members in education earn approximately the same as the average for women faculty at the university.

TABLE 2—PERCENTAGE SALARY INCREMENTS FOR MEN AND WOMEN
(Between Academic Ranks)

From	To	Male	Female
Teacher	Assistant Professor	25	25
Instructor	Assistant Professor	22	20
Assistant Professor	Associate Professor	21	21
Associate Professor	Professor	36	26

C. *Comparison of Men and Women's Salaries*

First, we discuss the Model I regression where sex is included as an independent variable. Its coefficient indicates that women earn, on average, 9.5 percent less than comparable men ($t=-7.0$); the pattern of the other coefficients remains about the same. A possible explanation of the differential is that employers use sex as a proxy for career commitment. In this case, the differential would be expected to decline, the longer the female employee remained with the university. Since time-series data are unavailable, we investigated this point by successively regressing the logarithm of income against the independent variables for all employees hired more than one, two, three, four, and finally, more than five years ago. The percentage differentials increase slightly. Thus there is no evidence to support decreasing differentials with increased length of employment.

Second, we compare some predicted salaries from the male- and female-regressions, using the same selected sets of characteristics. The person chosen arbitrarily as "standard" is a 35-year old, newly hired, white assistant professor in the "other" department with a Ph.D. A woman with these characteristics earns $12,510 which is 3.7 percent less than the corresponding man, who earns $12,990.[13] A 30-year old black fe-

---

[11] These coefficients are from an all-employee regression with both the age structure and department sex-specific.

[12] Bachelor degree holders who have achieved full professorship may well have had to "prove themselves" and accomplish more than those with the sinecure of a Ph.D.

[13] These percentages are found from the regression tables as follows. Let $\hat{z}_m$ and $\hat{z}_f$ be the predicted *log*

TABLE 3—SENSITIVITY ANALYSIS OF INCOME DIFFERENTIALS BY SEX[a]
(from standard)

	Percentage
Standard	−3.7
Age	
25	−7.2
35	−3.7
45	−8.9
55	−21.3
65	−19.9
Seniority	
0	−3.7
5	−2.6
Race	
Black	−10.6
White	−3.7
Education	
Bachelor's Degree	−15.4
Master's Degree	−4.2
Doctoral Degree	−3.7
M.D.	−13.6
Rank	
Instructor	−2.1
Assistant Professor	−3.7
Associate Professor	−3.5
Professor	−10.5
Department	
Fine Arts	−3.1
Engineering and Science	0.8
Social Science	7.4
Quant. Social Science	−6.9
Education	−5.8
Medicine	−4.8
Health	1.2
Other	−3.7

[a] The percentage differentials are derived from female salary minus male salary divided by male salary, each estimated from the appropriate logarithmic regression. For each entry, all characteristics other than the one being analyzed are the same as for the standard employee.

salaries from the male- and female-employee regressions and $\hat{\sigma}_m$ and $\hat{\sigma}_f$ be their standard errors of estimate. Under our assumption that salary follows a lognormal distribution, a predicted *log* salary of $\hat{z}_m$ yields a predicted dollar salary of $e^{\hat{z}_m+1/2\hat{\sigma}_m^2}$. Thus, the percentage difference in income (divided by 100) is given by

$$\frac{e^{\hat{z}_f+1/2\hat{\sigma}_f^2}-e^{\hat{z}_m+1/2\hat{\sigma}_m^2}}{e^{\hat{z}_m+1/2\hat{\sigma}_m^2}}=e^{\hat{z}_f-\hat{z}_m+1/2(\hat{\sigma}_f^2-\hat{\sigma}_m^2)}-1$$

Since $e^{1/2(\hat{\sigma}_f^2-\hat{\sigma}_m^2)}=1.0058$, ignoring this term biases the differential downward by 0.58 percent. Thus the interested reader may use the approximation $e^x-1=x$ and estimate the percentage differences in income for other cases by $(\hat{z}_f-\hat{z}_m)\cdot 100$. We present the exact values.

male instructor in education with a bachelor's degree and one year of seniority earns $10,280 or 23.1 percent less than a comparable man who earns $13,370. A 45-year old white female full professor in medicine with an M.D. and ten years seniority earns $35,260 or 23.0 percent less than a comparable man, who earns $45,790. On the other hand, a 35-year old white female assistant professor in the social sciences with a Ph.D. and three years seniority earns $14,980 or 8.1 percent more than a comparable man, who earns $13,860. A more systematic comparison between men and women is presented in Table 3 where the percentage differentials are tabled for the cases of each characteristic

TABLE 4—AVERAGE ACTUAL INCOME DIFFERENTIALS BY SEX[a]

	Number	Average Dollar Differential
All	323	−1,602
Age		
≤35	130	−722
35+−45	97	−1,116
45+−55	61	−3,264
55+	35	−3,320
Race		
Black	21	−2,014
White	295	−1,582
Education		
Bachelor's Degree	31	−1,797
Master's Degree	157	−924
Doctoral Degree	109	−1,535
M.D.	22	−6,605
Rank		
Instructor	115	−612
Assistant Professor	117	−1,306
Associate Professor	47	−2,091
Professor	22	−4,923
Department		
Fine Arts	33	−1,508
Engineering and Science	19	−726
Social Science	17	−1,000
Quant. Social Science	6	−3,067
Education	45	−1,902
Medicine	35	−4,594
Health	131	−1,002
Other	37	−1,403

[a] The percentage differentials are based on the women's actual salary minus that predicted for a comparable man from the logarithmic regression for men. Each differential is an average over all women in our sample with the given characteristic.

being varied, one at a time, from those of the standard person.

Finally, Table 4 gives the average dollar salary differential (between the woman's actual salary and that forecast for a comparable man from the male-regression) for the women actually employed at the university with the characteristic listed in the left-hand column. That is, each woman will enter one of the averages for age, one for education, etc. Care must be taken when interpreting these figures due to the interaction of various variables. For example, the \$6,600 average differential for a woman with a master's degree comes from a concentration of these women in ranks and departments with high percentage differentials. However, the figures do give information about the pattern of differentials at the university. The average (over all women faculty) is a \$1,602 differential (or 11.4 percent)[14] which is roughly comparable with the percentage differential from the all-employee regression of 9.5 percent.

V. Summary

The sources of wage differentials for the faculty of a large urban university were investigated. Between 1,000 and 2,000 faculty employees were included in the sample; sex, race, and department differentials received particular attention. Other independent variables included age, seniority, education, and academic rank. Two models were investigated, both using the logarithm of salary as the dependent variable, and a piecewise linear function for age. In Model I, sex was included as a dummy variable in a regression analysis of all employees. In Model II, which treated male and female employees separately, all coefficients were sex-specific.

A summary of the statistical results on the independent variables follows. Model I estimated that a woman on average earned 9.5 percent less than the corresponding male ($t=-7.0$). Using Model II, for each woman, the salary of a corresponding man was predicted (from the regression over male employees) and compared with the woman's actual salary. The average differential was -11.4 percent.

Model I predicted that a black would earn 13.0 percent more than a corresponding white ($t=5.5$). Model II estimated a black male to earn 14.0 percent more and a black female to earn 7.5 percent more ($t=5.1$, 1.9, respectively) than the corresponding whites.

Models I and II agree closely on the estimated differentials by department. Fine arts and "other" are the least well paid and about equal. Medicine is by far the highest paid. Social science *relatively* favors females, while education *relatively* favors males.

Seniority has significantly negative coefficients: -0.5 percent per year for all employees combined ($t=-5.4$), -0.6 percent for male employees only ($t=-5.7$), and -0.3 percent for female employees only ($t=-1.5$). The combined age profile peaks at about age 45. Each year of age is estimated to increase the employee's salary by 0.5 percent to 0.6 percent per year to age 45, but to decrease it by 0.1 percent to 0.2 percent per year thereafter. Sex-specific profiles are similar except that the female profile peaks about fifteen years earlier than that for men.

Education and rank variables behave as expected. A heavy premium is estimated for holders of an M.D. degree and for full professors. Profiles for men and women faculty considered separately are similar.

Finally, the adjusted R^2 for Model I was 0.79; the adjusted R^2 for Model II (male faculty) was 0.79, while for Model II (female faculty) it was 0.74.

---

[14] This percentage is obtained by dividing the \$1,602 differential by the average income the women would be predicted to have by the logarithmic regression for male employees.

REFERENCES

B. R. Bergmann, "Occupational Segregation, Wages and Profits When Employers Discriminate by Race or Sex," mimeo., Univ. Maryland, Jan. 1971.

M. S. Cohen, "Sex Differences in Compensation," *J. Hum. Resources*, Fall 1971, *6*, 434–47.

F. Y. Edgeworth, "Equal Pay to Men and Women for Equal Work," *Econ. J.*, Dec. 1922, *32*, 431–57.

E. Ginzberg, "Paycheck and Apron-Revolution in Woman Power," *Ind. Relat.*, May 1968, *7*, 193–203.

M. P. Goldberg, "The Economic Exploitation of Women," *Rev. Radical Polit. Econ.*, Spring 1970, *1*, 35–47.

N. M. Gordon and T. E. Morton, "A Low Mobility Model of Wage Discrimination—With Special Reference to Sex Differentials," *J. Econ. Theory*, forthcoming.

M. T. Hamilton, "A Study of Wage Discrimination by Sex: A Sample Survey in the Chicago Area," unpublished doctoral dissertation, Univ. Pennsylvania 1970.

J. N. Hedges, "Women Workers and Manpower Demands in the 1970's," *Mon. Lab. Rev.*, June 1970, *93*, 19–29.

M. S. Kassem, "The Salary Compression Problem," *Personnel J.*, Apr. 1971 313–17.

J. F. Madden, *The Economics of Sex Discrimination*, Durham 1972.

G. B. McNally, "Patterns of Female Labor Force Activity," *Ind. Relat.*, May 1968, *7*, 204–18.

R. L. Oaxaca, "Male-Female Wage Differentials in Urban Labor Markets," *Int. Econ. Rev.* forthcoming.

V. K. Oppenheimer, "The Sex Labeling of Jobs," *Ind. Relat.*, May 1968, *7*, 219–34.

C. D. Phelps, "Is the Household Obsolete?" *Amer. Econ. Rev. Proc.*, May 1972, *62*, 167–74.

H. Sanborn, "Pay Differences Between Men and Women," *Ind. Lab. Relat. Rev.*, July 1964, *17*, 534–50.

E. Waldman, "Changes in the Labor Force Activity of Women," *Mon. Lab. Rev.*, June 1970, *93*, 10–18.

F. B. Weisskoff, "'Women's Place' in the Labor Market," *Amer. Econ. Rev. Proc.*, May 1972, *62*, 161–66.

H. Wilensky, "Women's Work," *Ind. Relat.*, May 1968, *7*, 235–48.

H. Zellner, "Discrimination Against Women, Occupational Segregation, and the Relative Wage," *Amer. Econ. Rev. Proc.*, May 1972, *62*, 157–60.

Women's Bureau, Department of Labor, *Background Facts on Women Workers in the U.S.*, Washington 1970.

———, *Fact Sheet on the Relative Position of Women and Men Workers in the Economy*, Washington 1966.

———, *Report of the Task Force on Labor Standards*, Washington 1968.

———, *Unions and the Changing Status of Women Workers*, Washington 1964.

The Rationale of the Mean-Standard Deviation Analysis: Comment

By KARL BORCH*

Portfolio analysis based on mean and variance was introduced by Harry Markowitz (1952) more than twenty years ago. This method of analysis has become extremely popular, partly through Markowitz's own book (1959), and also through the work of James Tobin. The method has been severely criticized, and I am among the critics who have argued that mean-variance analysis must be seen as an intellectual exercise, useful only if it leads to insight into the real problems of portfolio management. The method has its staunch defenders, and one of the most articulate among these is S. C. Tsiang. In a recent article in this *Review*, he gives a comprehensive survey of the problem, and concludes that mean-variance analysis gives "a fair approximation—for most practical purposes."

I have no objection to this cautious conclusion, although the term "practical purposes" may be given unwarranted interpretations. Tsiang does, however, devote a considerable amount of space to a discussion of a simple classroom example which I published a few years ago, and it may be in order to spell out the ideas behind the example. It is convenient first to present the example again in a less general form than the original. Let us consider the following two probability distributions or "gambles":

Gamble A,	which will give:
	0 with probability 0.5
or	2 with probability 0.5
Here:	Mean $=1$ and Variance $=1$

Gamble B,	which will give:
	-2 with probability 0.2
or	3 with probability 0.8
Here:	Mean $=2$ and Variance $=4$

I will not question the rationality of a person who considers the two gambles as equally attractive. If, however, he says that the gambles are equivalent *because* the mean-variance pairs $\{1, 1\}$ and $\{2, 4\}$ are equivalent, I will present him with:

Gamble C,	which will give:
	0 with probability 0.5
	4 with probability 0.5

This gamble is clearly more attractive than A, and hence it should also be more attractive than B. The gambles B and C do, however, both have the mean-variance pair $\{2, 4\}$, and this should imply that they are equivalent.

For any two gambles of this form, alleged to be equivalent, we can construct a third gamble of the same form which:

1) is clearly superior to one of the original gambles
2) has the same mean-variance pair as the other original gamble

This proves that a consistent preference ordering over the set of all probability distributions cannot be represented by a function of mean and variance, i.e., a utility function of the type $U(E, V)$.

I believe that Tsiang accepts this conclusion, and that he will take the example as an illustration that a mean-variance representation may not always give a fair approximation. He argues that in such cases we have to consider the skewness of the two distributions. Gamble B has a negative skewness, and hence it must be less attractive than the symmetric gamble C. This argument just passes the buck to authors with more patience for elaborate arithmetic. It is easy to escalate the example. For two gambles alleged to be equivalent, we can compute the mean-variance-skewness triples, and then construct a third gamble with the

---

* Norwegian School of Economics and Business Administration.

same triple as one of the original games, and clearly superior to the other one.

I shall not take the reader through the arithmetic involved in such escalating. Instead I shall present another classroom example, which may give some further insight into mean-variance analysis. The *gamma* density

$$g_n(x) = \frac{\alpha^{n+1}}{n!}(x-c)^n e^{-\alpha(x-c)} \qquad c \leq x$$

has the mean $E=(n+1/\alpha)+c$ and the variance $V=n+1/\alpha^2$.

Let us consider two points in the EV plane, (E_1, V_1) and (E_2, V_2) which are alleged to be on the same indifference curve. We can then construct two *gamma* distributions $g_{n_1}(x)$ and $g_{n_2}(x)$, corresponding to these points, by taking:

$$c = \frac{E_1V_2 - E_2V_1}{V_2 - V_1}$$

$$\alpha = \frac{E_2 - E_1}{V_2 - V_1}$$

$$n_1 = \left(\frac{E_2 - E_1}{V_2 - V_1}\right)^2 V_1 - 1$$

$$n_2 = \left(\frac{E_2 - E_1}{V_2 - V_1}\right)^2 V_2 - 1$$

These two distributions cannot possibly represent equivalent gambles. If $n_2 > n_1$, we have

$$\text{(1)} \qquad \int_c^z g_{n_1}(x)dx > \int_c^z g_{n_2}(x)dx \qquad \text{for all } z \text{ in } (c, \infty)$$

This implies that $g_{n_2}(x)$ is superior to $g_{n_1}(x)$. For any z, $g_{n_2}(x)$ offers the greater probability of a gain exceeding z. If n_1 and n_2 are integers, (1) follows from straightforward evaluation of the integrals. For a more general proof, consider:

$$\int_c^z \{g_{n_1}(x) - g_{n_2}(x)\}dx$$

$$= \int_c^z \left\{\frac{\alpha^{n_1+1}}{n_1!}(x-c)^{n_1} - \frac{\alpha^{n_2+1}}{n_2!}(x-c)^{n_2}\right\} \cdot e^{-\alpha(x-c)}dx$$

$$= \int_0^y \left\{\frac{x^{n_1}}{n_1!} - \frac{x^{n_2}}{n_2!}\right\} e^{-x}dx = F(y)$$

We have $F(o)=F(\infty)=0$, and further

$$F'(y) = \frac{y^{n_1}}{n_1!}\left\{1 - \frac{n_1!}{n_2!}y^{n_2-n_1}\right\}e^{-y}$$

Hence with increasing y, $F(y)$ will increase from 0 to a maximum, and then decrease to zero as y goes to infinity. This implies that $F(y)>0$ for any finite, positive y.

It is worth noting that the skewness of $g_n(x)$ is $2(n+1)^{-\frac{1}{2}}$. Hence with increasing n, the skewness will decrease, whilst $g_n(x)$ itself becomes more attractive. Tsiang's conclusion that skewness is an attractive property can not be generally valid.

It may be worthwhile trying to clarify a question which frequently has confused discussions about the validity of the mean-variance approach to portfolio analysis. Let us consider a set of stochastic variables with distributions belonging to a family determined by two parameters, say $F(x, \alpha, \beta)$. There will be a one-to-one correspondence between the two parameters and the mean and the variance, except in degenerate cases. We will, therefore, lose little by writing $F(x, E, V)$. A preference ordering over this set of stochastic variables will be a preference ordering over the set of ordered pairs (E, V), and it can be represented by a set of indifference curves in an EV plane. Assume now that x_1 and x_2 belong to the set under consideration, and that their distributions are $F(x, E_1, V_1)$ and $F(x, E_2, V_2)$, respectively. A portfolio of these two "assets" is then defined by a stochastic variable $z=kx_1+(1-k)x_2$, with $0<k<1$. The mean and the variance of z are:

$$E = kE_1 + (1-k)E_2$$

$$V = k^2V_1 + (1-k)^2V_2$$

The distribution of z will not in general be $F(x, E, V)$, but a distribution of a different type, say $G(x, E, V)$. This distribution will have no place in our original preference ordering. If we give it a place by requiring that $G(x, E, V)$ shall be equivalent to $F(x, E, V)$ for any E and V, we will run into contradictions of the kind illustrated by the examples given above. G will be of the same type as F, only if F belongs to the class of distributions called "stable." The only member of this class, which has a finite variance, is the normal distribution. Hence portfolio analysis based on mean-variance methods can be applied without restriction only when all assets are represented by normally distributed variates. It may also be worthwhile recalling that a probability distribution is not uniquely determined by its moments. A textbook example (see William Feller, p. 224) is

$$(2)\quad f(x) = \frac{1}{24}\left\{1 - \alpha \sin \sqrt[4]{x}\right\} \exp\left(-\sqrt[4]{x}\right)$$

which represents a probability density over $(0, \infty)$ if $0<\alpha<1$. The nth moment is $m_n = (4n+3)!/6$ which is independent of α. Hence, if we seek a preference ordering based on mean, variance, skewness and higher moments, we start with a behavioral assumption to the effect that all distributions of the form (2) are considered as equivalent. This assumption can be rejected out of hand, and so can mean-variance analysis. In spite of this, I shall continue to use mean-variance analysis in teaching, but I shall warn students that such analysis must not be taken seriously and applied in practice.

REFERENCES

K. **Borch**, "A Note on Uncertainty and Indifference Curves," *Rev. Econ. Stud.*, Jan. 1969, *36*, 1–4.

W. **Feller**, *An Introduction to Probability Theory and Its Applications*, Vol. II, New York 1966.

H. M. **Markowitz**, "Portfolio Selection," *J. Finance*, Mar. 1952, *7*, 77–91.

———, *Portfolio Selection—Efficient Diversification of Investments*, New York 1959.

J. **Tobin**, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, *25*, 65–85.

S. C. **Tsiang**, "The Rationale of the Mean-Standard Deviation Analysis, Skewness Preference, and the Demand for Money," *Amer. Econ. Rev.*, June 1972, *62*, 354–71.

The Rationale of the Mean-Standard Deviation Analysis: Comment

By G. O. Bierwag*

In a recent paper in this *Review*, S. C. Tsiang severely criticizes the traditional portfolio analysis within the $(E\text{-}S)$ framework. Tsiang propounds two major contentions: the $(E\text{-}S)$ framework is appropriate 1) only when E is large relative to S, and 2) when the indifference curves in the $(E\text{-}S)$ half-plane have slopes of 45° or less. I have no quibble with the first of these conditions; it is relatively well-founded in the work of Kenneth Arrow, John Pratt, and Paul Samuelson and it is based on reliable principles of approximation. Tsiang, using the second contention, shows that the $(E\text{-}S)$ framework cannot support James Tobin's theory of liquidity preference. This would, indeed, be a serious defect of the $(E\text{-}S)$ analysis, if it were true. Fortunately, the second contention is quite false. In this comment, I show why this is so and I propose a substitute contention that meets much of the spirit of Tsiang's inquiry.

I

Consider an example in which an investor may invest in two assets—one riskless and the other risky. Let x_0 and x_1 be the number of dollars, respectively, so invested in each. At the end of the investment period, the initial investment of x_0+x_1 dollars grows to $x_0+\lambda x_1=R$ dollars where it is supposed that λ is a random variable with a subjective cumulative probability distribution, $F(\lambda)$. Assume that $F(\lambda)$ is well specified by two independent parameters, μ and σ, the mean and standard deviation of λ. Let $u(R)$ be a twice differentiable utility function over wealth with the property that $u'(R)>0$ and $u''(R)<0$, i.e., the investor is risk averse. Under the above assumptions, expected utility can be expressed as a function of only the mean and standard deviation of the portfolio. Letting $E=x_0+\mu x_1$ and $S=\sigma x_1$, one may write

$$(1)\qquad U(E, S) = E[u(R)]$$

The investor now faces the problem of choosing E and S by appropriately choosing x_0 and x_1. This choice is equivalent to choosing some distribution of R, subject to the fixed distribution of λ. As Myron Grove and the author have noted and developed, the expected utility in (1) has contours of constant expected utility which map into the (x_0, x_1) space. Naturally, each point on a contour in that space maps into a point on the corresponding contour in the $(E\text{-}S)$ half-plane. Consider the slope of an indifference curve in the $(E\text{-}S)$ half-plane. Taking the total derivative of $U(E, S)$ with respect to x_1, one has

$$(2)\qquad U_E\left[\frac{dx_0}{dx_1}+\mu\right]+U_S\sigma = 0$$

along a contour of constant expected utility where dx_0/dx_1 is the corresponding slope of the contour in the (x_0, x_1) space. Then, the slope of an indifference curve in the $(E\text{-}S)$ space is

$$(3)\qquad \frac{dE}{dS} = -U_S/U_E = \frac{\mu}{\sigma}+\frac{1}{\sigma}\frac{dx_0}{dx_1}$$

Now, if μ is large relative to σ, it is highly likely that dE/dS exceeds unity; but, of course, everything depends on dx_0/dx_1 which we should naturally expect to be negative because x_0 and x_1 are to some extent substitutable investments. However, one can easily illustrate with examples that dx_0/dx_1 need not be so negative as to imply that $dE/dS\leq 1$. For example, suppose that λ is *gamma* distributed and that $u(R)=-\exp(-R)$. The expected utility is then

$$(4)\qquad E[u(R)] = \frac{-1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty \lambda^{\alpha-1}\cdot\exp(-x_0-\lambda x_1-\lambda/\beta)d\lambda$$

---

* Professor of economics, University of Oregon.

where α and β are positive parameters of the distribution of λ. Resolving the integral in (4),

$$(5)\quad E[u(R)] = -(1+\beta x_1)^{-\alpha}\exp(-x_0)$$

Holding expected utility fixed at some level and allowing variations in x_0 and x_1, it follows that

$$(6)\quad (1+\beta x_1)dx_0 + \alpha\beta dx_1 = 0$$

along an indifference curve in the (x_0, x_1) space. The slope of the indifference surface is then

$$(7)\quad \frac{dx_0}{dx_1} = \frac{-\alpha\beta}{1+\beta x_1}$$

Noting that $\beta=\sigma^2/\mu$ and $\alpha=\mu^2/\sigma^2$, (7) becomes

$$(8)\quad \frac{dx_0}{dx_1} = \frac{-\mu^2/\sigma^2}{x_1+\dfrac{\mu}{\sigma^2}}$$

The slope is very small in absolute value when x_1 tends to be large. Referring to (3), it is clearly evident that the slope, dE/dS, may exceed unity.[1]

The problem with Tsiang's approach is that he doesn't maintain the constancy of the parameters μ and σ. Evidently the subjective distribution of λ depends on the investor's choice of x_0 and x_1. Why this should be so is not clear. He appears to be assuming that if λ is distributed according to a lognormal distribution, say, then so is R. Clearly, this is not so.

Tsiang's 45° rule was prompted by an effort to avoid Karl Borch's problem in which Bernouilli mixtures of distributions were relevant choices in an asset space. The point is that different choices of assets in portfolio theory are *not* also choices of Bernouilli mixtures.[2]

II

In portfolio theory, as Tsiang suggests, it seems reasonable to suppose that Prob $[\lambda<0]=0$ This assumption reflects the fact that many financial instruments have provisions for limited liability; there are stop-loss provisions in stock and commodity markets, etc. Suppose that λ, then, is so distributed. In the $(E\text{-}S)$ framework, this restriction on λ must imply that the indifference curves have an elasticity which is less than unity. The proof may proceed as follows. The marginal expected utilities for the assets must be positive.[3] That is,

$$(9)\quad E[u'(R)] = U_E > 0$$

and

$$E[u'(R)\lambda] = U_E\mu + (U_S/S)\sigma^2 x_1 > 0$$

Multiplying the first of these inequalities by x_0 and the second by x_1, so that in adding the equations together, one has

$$(10)\quad U_E E + U_S S > 0$$

from which it is clear that

$$(11)\quad \frac{dE}{dS}\,\frac{S}{E} = -(U_S/U_E)(S/E) < 1$$

---

[1] This choice of the *gamma* function is strictly for convenience. Almost any two-parameter distribution of λ works satisfactorily as long as expected utility exists. The utility function, $-\exp(-R)$, is a bounded function for $R\geq 0$, and the measure of relative risk aversion is simply R which is unbounded from above. Given this situation, one can use one of Arrow's theorems (1971) to show that the investor will always hold money with a sufficiently large initial wealth—and this is so regardless of how many independent parameters are in the probability distribution function.

[2] The von Neumann-Morgenstern axioms still hold for this situation. Although these axioms apply to "mixture sets," there is absolutely no reason to suppose that *every* problem of choice is one that requires the points of choice to be mixtures of probability distributions. The set of portfolio choices is a *subset* of an entire mixture set to which the von Neumann-Morgenstern axioms apply. Borch's problem and Tsiang's refinement of it are consequently irrelevant to the problem of portfolio selection as it is traditionally stated. Other problems of choice under uncertainty may well involve mixtures and Borch's problem may well apply, but this is not so for the portfolio theory that gives rise to a theory of liquidity preference.

[3] Grove and I (1966) consider the case in which there is no riskless asset in the portfolio. If one allows Prob$[\lambda<0]>0$ for each asset, it is possible to get cases in which all assets have negative marginal utilities. Such a portfolio is then "contaminated by risk." The above assumptions thus prevent contamination.

provided the investor is risk averse, $x_0>0$, and $x_1>0$. If one, therefore, assumes that a portfolio can never be contaminated, then (11) must hold. Of course, Tobin's theory of liquidity preference can still be maintained under this restriction.

REFERENCES

K. J. Arrow, *Aspects of the Theory of Risk-Bearing,* Helsinki 1964.

———, *Essays in the Theory of Risk-Bearing,* Chicago 1971.

G. O. Bierwag and M. A. Grove, "Indifference Curves in Asset Analysis," *Econ. J.,* June 1966, *76,* 337–43.

——— and ———, "Slutsky Equations for Assets," *J. Polit. Econ.,* Jan./Feb. 1968, *76,* 114–27.

K. Borch, "A Note on Uncertainty and Indifference Curves," *Rev. Econ. Stud.,* Jan. 1969, *36,* 1–4.

J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica,* Jan. 1964, *32,* 122–36.

P. A. Samuelson, "The Fundamental Approximation Theorem of Portfolio Analysis in Terms of Means, Variances, and Higher Moments," *Rev. Econ. Stud.,* Oct. 1970, *37,* 537–42.

S. C. Tsiang, "The Rationale of the Mean-Standard Deviation Analysis, Skewness Preference, and the Demand for Money," *Amer. Econ. Rev.,* June 1972, *62,* 354–71.

J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.,* Feb. 1958, *25,* 65–85.

The Rationale of the Mean-Standard Deviation Analysis: Comment

By HAIM LEVY*

In a recent article in this *Review*, S. C. Tsiang examines the rationale of the mean-variance (or mean-standard deviation) analysis. Accepting his findings might lead one to some definite conclusions regarding the importance and the economic significance of the mean-variance analysis. At first sight, it seems that Tsiang's paper presents the following encouraging results:

1) It resolves the paradoxes raised by Karl Borch and Martin Feldstein.
2) It shows that the slope of the indifference curves is almost always less than 1.
3) It maintains that the mean-variance rule can serve as a good approximation rule for a wide range of utility functions.
4) It shows that the demand to include cash assets in the investment portfolio is not justifiable even where the existence of riskless bonds is ruled out.

It is the intention of this paper to show that even if Tsiang's assumptions for a good approximation are valid, one cannot apply the mean-variance analysis while ignoring all higher distribution's moments. Moreover, the slope of the *relevant* indifference curves is not in general less than 1; hence the paradoxes presented by Borch and Feldstein cannot be resolved according to Tsiang's suggestion, and his criticism of James Tobin's demand for cash analysis is invalid. Finally, it is shown that indifference curves which are defined in terms of the relevant parameters are in fact convex even in the case of log-normal distribution of returns. A distinction must be made between a *sufficient* investment rule and an *optimal* investment rule (which is a sufficient as well as a necessary rule) in order to explain Borch-Feldstein's paradoxes.

I. The Mean-Standard Deviation Criterion and Higher Distribution Moments

Expanding the utility function $U(y)$ into Taylor's series, Tsiang arrives at:

$$(1)\quad U(y) = U(\bar{y}) + U'(\bar{y})h + U''(\bar{y})\frac{h^2}{2!} + U'''(\bar{y})\frac{h^3}{3!} + \ldots + R_n$$

where y is the investor's wealth which is a random variable, $\bar{y}$ is the expected wealth, and R_n is a remainder term.

Taking the expected utility of both sides of (1) yields

$$(2)\quad EU(y) = U(\bar{y}) + U''(\bar{y})\frac{\bar{m}_2}{2!} + U'''(\bar{y})\frac{\bar{m}_3}{3!} + \ldots + ER_n$$

where $\bar{m}_i$ is the ith central moment of the random variable.

Where the above series is convergent, R_n can be ignored. Clearly, in a case where a quadratic utility function is assumed, $EU(y)$ is determined only by the first two terms of equation (2) since all other derivatives are zero. However, Tsiang maintains that the higher moments can be ignored not only in the case of a quadratic utility function (or normal distribution of returns), but under certain conditions, also with other important utility functions. Let us start with the negative exponential function $U(y) = B(1-e^{-\alpha y})$. Expanding this function into Taylor's series, he arrives at:

---

* The Hebrew University, Jerusalem. I wish to thank Jerome Stein and a referee for their helpful comments.

$$(3)\quad E[U(y)] = B - Be^{-\alpha\bar{y}}\left[1 + \frac{\alpha^2S^2}{2!} - \frac{\alpha^3\bar{m}3}{3!} + \frac{\alpha^4\bar{m}4}{4!} - \dots\right]$$

The above equation can be rewritten as:

$$(4)\quad E[U(y)] = B - Be^{-\alpha\bar{y}}\left[1 + \frac{\alpha^2S^2}{2!} - \frac{\alpha^3S^3}{3!}\mu_3 + \frac{\alpha^4S^4}{4!}\mu_4 - \dots\right]$$

where μ_3 is the pure number measure for skewness, i.e., $\mu_3 = \bar{m}_3/S^3$ and similarly $\mu_4 = \bar{m}_4/S^4$, S being the standard deviation. Tsiang claims that it is reasonable to assume that $\alpha S < 1$. I will not argue with this assumption and, for the time being, accept it. However, even if $\alpha S < 1$, one cannot claim that the mean-standard deviation analysis (M-S) provides a good approximation and that higher moments can be ignored. Obviously, the coefficients of μ_3 and μ_4 are, as Tsiang claims, very small. But μ_3 might be large and the mistake incurred by neglecting S^2 might be *smaller* than the mistake resulting from ignoring $\bar{m}_3$.

A simple numerical example will clarify this point. Suppose that a distribution of y is given by

$P(y)$	y
1/100	100
99/100	0

and that $\alpha S = 1/2$ (i.e., smaller than suggested by Tsiang). In this example, μ_3 is about 10, and hence

$$\frac{\alpha^2S^2}{2!} = 0.125$$

$$\frac{\alpha^3S^3}{3!}\mu_3 = 0.208$$

When we consider equation (4), it becomes clear that by ignoring the variance rather than the skewness, the mathematical error is reduced and a better approximation is achieved. Since μ_3, μ_4, etc. are not limited in their absolute values, there is no economic sense in ignoring higher moments even if $\alpha S < 1$. Hence there is no guarantee that the M-S analysis will provide a good approximation even for the negative exponential utility function.

As regards the important utility functions $U(y) = log\ (y)$ and $U(y) = (1/(1-\alpha))y^{1-\alpha}$, $(\alpha > 0)$, Tsiang maintains that the M-S analysis provides a good approximation to the expected utility criterion if $S/\bar{y}$ is relatively small, say, smaller than 1/10. However, if $S/\bar{y}$ is large, the M-S analysis might lead to paradoxical results because the coefficient of the distribution's skewness, not being small, cannot be ignored. More specifically, while $U(y) = log\ (y)$, $EU(y)$ can be written as

$$EU(y) = log\,\bar{y} - \frac{1}{2}\frac{S^2}{\bar{y}^2} + \frac{1}{3}\frac{S^3}{\bar{y}^3}\mu_3 - \frac{1}{4}\frac{S^4}{\bar{y}^4}\mu_4 + \dots$$

At first sight it may seem that while $S/\bar{y} < 1$, higher moments can be ignored. Let us demonstrate that this conclusion is false by using the following two rectangular distributions:[1]

$$y_1 \sim R(98.27,\ 98.27 + \sqrt{12})$$
$$y_2 \sim R(196.535,\ 196.535 + \sqrt{48})$$

The expected value and standard deviations of the above two distributions are given by

		$\bar{y}$	S
y_1	:	100	1
y_2	:	200	2

Thus $S/\bar{y} = 1/100$, which is very small. According to Tsiang, the M-S analysis is in this case valid for a wide family of utility functions. Since $E(y_2) > E(y_1)$, $S(y_2) > S(y_1)$, according to the M-S criterion, y_1 and y_2 should be included in the efficient set and thus some risk averters are "approximately indifferent" between y_1 and y_2. Note that y_1 and y_2 are symmetric and hence μ_3 can be ignored. The coefficient of μ_4 is $S^4/4\bar{y}^4 = 1/4 \times 10^8$, which is, of course, very small. Thus according to Tsiang's argument, all higher moments can be ignored and one can safely use the M-S criterion in the choice between y_1 and y_2.

---

[1] The Rectangular distribution is a two-parameter (symmetric) distribution. If $y \sim R(\alpha, \beta)$ then the density function is given by $f(y) = 1/(\beta - \alpha)$, and $Ey = (\alpha + \beta)/2$, $\sigma_y^2 = (\beta - \alpha)^2/12$.

However, even if the coefficients of the higher moments are indeed very small, one cannot ignore them and the M-S analysis cited is invalid. Another look at the above example makes it clear that every rational investor would choose y_2 rather than y_1. There is no need even to assume risk aversion in order to reach this conclusion, since for *every* utility function U with $U' > 0$, $EU(y_2) > EU(y_1)$ since $F(y_2) < F(y_1)$, where F denotes a cumulative probability distribution.[2] Indeed for the utility function $U(y) = \log y$, $EU(y_1) = 4.3$, $EU(y_2) = 5.2$. Thus as one would assume, y_2 yields a higher expected utility than y_1. The above instance makes it clear that similar examples can be cited with $S/\bar{y} < \epsilon$, where ϵ may be as small as one wishes. Nevertheless, contrary to Tsiang's conclusion, the M-S analysis does not provide a good approximation.

II. Borch's Paradox and the Slopes of Indifference Curves

Borch maintains that it is generally impossible to draw a family of indifference curves in the M-S space. He presents the following two options: (x_0, p, x_1) and (x_0, p, x_2); namely there is a probability p to get x_0, and a probability $(1-p)$ to get x_1 under the first option, and x_2 under the second option. Assuming that $x_1 > x_2$, it is clear that all investors would prefer the option (x_0, p, x_1) to (x_0, p, x_2). However, the expected return and the standard deviation of the first option are respectively larger than the expected return and the standard deviation of the second option. Namely $E_1 > E_2$, $S_1 > S_2$ and on an indifference map drawn according to the M-S criterion, the two options lie on an appropriate indifference curve. Since it is clear that the first option dominates the second one, Borch concludes that it is generally impossible to draw the M-S indifference curves. Tsiang suggests a solution to this paradox: he maintains, and in some instances proves, that the slope of the M-S indifference curve must be smaller than 1. Since the probability p is given by

$$p = \frac{(E_2 - E_1)^2}{(E_2 - E_1)^2 + (S_2 - S_1)^2}$$

and $(E_2 - E_1)/(S_2 - S_1) < 1$ (assuming that the slope is smaller than 1) he concludes that $p < 1/2$, and thus the two options presented by Borch must be positively skewed. Moreover, the distribution (x_0, p, x_1) is positively skewed to a greater extent than the distribution (x_0, p, x_2). Neglect of the third moment in Borch's M-S analysis leads to an underestimation of the value of option (x_0, p, x_1) in comparison with option (x_0, p, x_2). Thus, M-S analysis indicates a lack of indifference between the two options in spite of the fact that (x_0, p, x_1) would dominate (x_0, p, x_2) if the third moment were taken into account.[3]

In order to demonstrate that Tsiang's resolution of the paradox is illogical, let us analyze an example which in its essentials is similar to the example given by Borch. Let us examine the following two rectangular distributions

$$y_1 \sim R(9, 11)$$
$$y_2 \sim R(9 \times 10^6, 11 \times 10^6)$$

Since $Ey_2 > Ey_1$, $S(y_2) > S(y_1)$, on an indifference map drawn according to the M-S analysis the two options lie on an appropriate indifference curve. Since the two distributions are symmetrical, a neglect of the skewness in the M-S analysis does not result in any bias. Moreover, $S/\bar{y}$ is very small[4] and, according to Tsiang, use of the M-S is in this case more or less justifiable. However, no rational decision maker would say that he is indifferent or even more or less indifferent to the choice represented by the above two options! Thus, while $S/\bar{y}$ is small and skewness equal to zero, the M-S analysis nevertheless leads to paradoxical results.

---

[2] For proof of this statement see James Quirk and Rubin Saposnik, Josef Hadar and W. R. Russell, and Giora Hanoch and the author (1969).

[3] However, Tsiang maintains that if the standard deviation is small relative to total wealth, then the treatment of the above two options as approximately indifferent is also justifiable, see p. 363.

[4] Since $\bar{y} = (\alpha + \beta)/2$ and $S = (\beta - \alpha)/\sqrt{12}$, we get $S/\bar{y} = (11-9)/(9+11)/2\sqrt{12} = 0.06$ for the first distribution. Obviously $S/\bar{y}$ for the second distribution is also equal to 0.06.

Tsiang maintains that indifference curves in the M-S space will generally have a slope smaller than 1. Although this is true in the case of certain specific utility functions, in general this is incorrect. Let us start with the quadratic utility function $U(y)=y+by^2$ (where $b<0$). Tsiang maintains that along a given indifference curve $d\bar{y}/dS<1$. Differentiating $U(y)$ and establishing $dEU(y)=0$, one gets

$$d\bar{y}/dS = \frac{S}{-\dfrac{1}{2b} - \bar{y}}$$

Since $U'(y)=1+2by>0$, it is clear that $y<-1/2b>0$ and hence also $\bar{y}<-1/2b$. Thus $d\bar{y}/dS>0$. However, from the above equation it is not clear why $d\bar{y}/dS<1$ in every case, as claimed by Tsiang. One may choose any possible set of numbers for S, $\bar{y}$, and b, and without contradicting the basic assumptions ($U'(y)>0$; $U''(y)<0$) arrive at the formula that $d\bar{y}/dS>1$. Tsiang supports his claim that $d\bar{y}/dS<1$ with the following argument: the expected utility of a quadratic utility can be written as $E[U(y)]=U(\bar{y})+U''\ S^2/2$, hence $d\bar{y}/dS=-SU''/U'$, (for $E(y)=$ constant). *If* $d\bar{y}/dS>1$ one finds that $U'(\bar{y})+SU''(\bar{y})<0$.

Tsiang states:

> . . . which means that, provided the quadratic approximation is a close approximation within this range, the marginal utility of wealth would be brought to zero or to a negative value by a mere deviation of one standard deviation of the actual value of wealth from its mean. [p. 364]

The possibility that the deviation of one standard deviation from $\bar{y}$ will reduce the marginal utility to zero (or negative) is not at all surprising when dealing with a quadratic utility function. It is well known that the function $U(y)=y+by^2$ ($b<0$) with $U'(y)=1+2by>0$ and $U''(y)=2b<0$ is limited only to the range $y<-1/2b$. However, if $y\equiv\bar{y}+S$ happens to be greater than $1/-2b$, $U'(y)$ is negative. This is one of the defects of the quadratic utility function and one cannot use this argument in order to prove that the slope of the indifference curves is smaller than 1.[5]

Most of Tsiang's conclusions are based on the assumption that the slope of the indifference curves in the M-S space is smaller than 1. He shows that for the utility function $U(y)=B(1-e^{-ay})$ and where the returns are normally distributed, the slope of the indifference curve is equal to αS (S being the standard deviation). It is not apparent why αS must in general be smaller than 1. But even if αS is indeed smaller than 1, this does not provide a basis for Tsiang's far-reaching conclusion that the slope of the indifference curves must be smaller than 1. His criticism of Tobin's liquidity preference is also not justified since it is based on the assumption that the slope of the indifference curves in the M-S space is smaller than 1. The following simple example shows that, in general, the slope of the indifference curve is not limited at all, and therefore all Tsiang's conclusions that stem from this assumption are in doubt. Let y_1 and y_2 represent two alternative options with rectangular distributions, where

$$y_1 \sim R(100,\ 100 + 10\sqrt{12}),$$
$$y_2 \sim R(101,\ 101 + 2\sqrt{12})$$

The expected return and the standard deviation of the two options are given by: $\bar{y}_1=100+5\sqrt{12}$, $S_1=10$, and $\bar{y}_2=101+\sqrt{12}$, $S_2=2$. Since the two cumulative distributions of y_1

---

[5] This is easy to demonstrate with a numerical example, for which $U'(y)>0$, $U''(y)<0$, $y<-1/2b$ as required; nevertheless the deviation of one standard deviation from $\bar{y}$ (i.e., $y=\bar{y}+S$), will bring us to the declining part of the utility function, i.e., $U'(y)<0$, for $y=\bar{y}+S$. Let us choose the following random variable:

Probability	Outcome
1/10	$0=y_1$
9/10	$100=y_2$

Hence $\bar{y}=90$ and $S=30$. Let $b=-1/204$, hence max $(y)=100<-1/2b=102$ as required by the constraint $U'(y)>0$. Thus, y is restricted to the positive part and all the requirements are met. However, for the point $y=\bar{y}+S=120$ $U'(y)<0$.

and y_2 intersect, so that

$$\int_{-\infty}^{101} [F(y_2) - F(y_1)]dy < 0$$

$$\text{and} \int_{-\infty}^{\infty} [F(y_1) - F(y_2)]dy < 0$$

one may conclude that y_1 as well as y_2 should be included in the risk averter's efficient set.[6] Namely, there is one concave utility function U_1 for which $EU_1(y_1) > EU_1(y_2)$, and there is another concave utility function, U_2, for which $EU_2(y_1) < EU_2(y_2)$. Assuming that the two options y_1 and y_2 are mutually exclusive, the *M-S* criterion is an optimal rule.[7] There is thus an indifference curve which passes through y_2 and rises above y_1. For other risk averters the opposite is true. These two possible indifference curves are drawn in Figure 1. Since the slope of the straight line connecting y_2 with y_1 is greater than 1,

$$\frac{\bar{y}_1 - \bar{y}_2}{S_1 - S_2} = \frac{100 + 5\sqrt{12} - 101 - \sqrt{12}}{10 - 2} = \frac{4\sqrt{12} - 1}{8} > 1$$

it is clear that the slope of the possible indifference curve *II* is also greater than 1, thus contradicting Tsiang's assumption.[8]

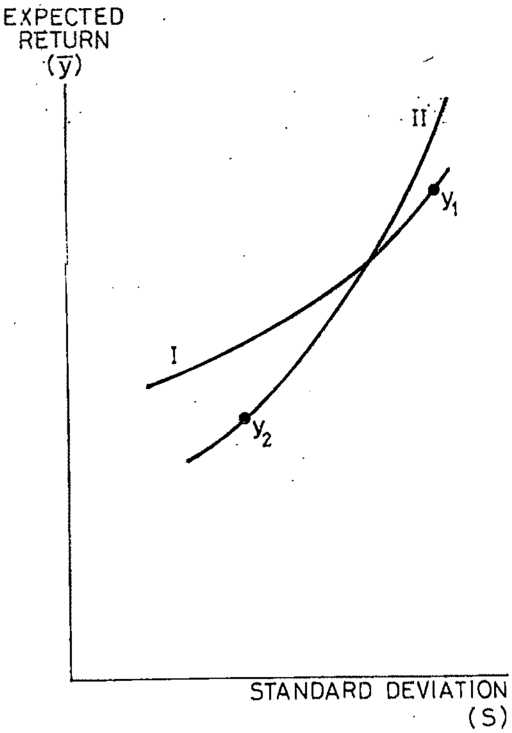


FIGURE 1

III. Feldstein's Paradox and the Convexity of the Indifference Curves

In this section we discuss the properties of the mean-variance rule in the case of log-normal distribution of returns and risk aversion. Feldstein has shown that for the log-normal distribution the Markowitz-Tobin mean-variance rule is invalid. Tsiang claims that Feldstein's paradox results from neglecting the positive skewness of the lognormal distributions. Moreover, Tsiang demonstrates that the slope of Feldstein's indifference curves is smaller than 1 which confirms his general conclusion. Feldstein's indifference curves are characterized by a concave as well as a convex segment. Tsiang suggests that we need consider only the convex part, which is, of course, consistent with Tobin's analysis.

Tobin, in his (1969) reply to Borch and Feldstein, maintains that the *M-S* indifference curves do not in general exist as loci of constant expected utility. Thus, the basic question inherent in the Borch-Feldstein-Tobin dispute is: when is it permissible to draw indifference curves in the *M-S* space, and when does the slope of the indifference curves have economic significance?

In order to analyze this problem, let us assume a two-parameter investment criterion which we will call the (Q_1, Q_2) criterion. This efficiency criterion which provides a partial ordering of uncertain prospects might be arbitrary and without economic significance, it might be a sufficient criterion for partial ordering, and it might be a necessary and sufficient criterion. A sufficient criterion is not, of course, an optimal criterion.[9] When using a sufficient criterion for a first screening of portfolios, one does not use all the available information, and therefore the Markowitz-Tobin efficient set might be too large, i.e., it may contain some portfolios that should have been eliminated from it, and this may result in paradoxes similar to those presented by Borch and Feldstein. For example, the *M-S* criterion is no more than a sufficient criterion for the quadratic utility function (see Hanoch and the author (1970)). Therefore, the use of the *M-S* rule with a quadratic utility function may create paradoxical results. Similarly, the *M-S* rule is no more than a sufficient rule for the set of options given by Borch. Thus, the efficient set is too large and includes (x_0, p, x_1) and (x_1, p, x_2) in spite of the fact that we know that one of these options should be excluded from it. We shall show below that the *M-S* rule is also no more than a sufficient rule for the lognormal distributions, thus resulting in the paradox of the nonconvex indifference curves arrived at by Feldstein.

---

[6] For further details see Hadar and Russell, Hanoch and the author (1969), and Michael Rothschild and Joseph Stiglitz.

[7] See Hanoch and the author (1969).

[8] Another even simpler example confirms the above conclusion. Let $y_1 \sim N(5, 6)$, $y_2 \sim N(15, 7)$. Since *M-S* rule is optimal for the above normal distributions, it is clear that $EU_1(y_1) > EU_1(y_2)$ for some concave U_1 and the opposite is true for another concave U_2. However, $\bar{y}_2 - \bar{y}_1/(S_2 - S_1) = 10$ which, of course, implies that some indifference curves exist with a slope which is greater than 1.

Finally, if (Q_1, Q_2) is an optimal rule for a partial ordering of portfolios (i.e., sufficient and necessary) one would expect it to lead, in general, not to paradoxes but to well-behaved, convex, indifference curves.

Since the lognormal distribution is economically interesting,[10] we may here analyze it in greater detail. We show below that choosing $Q_1 = Ex$, $Q_1 = \sigma(x)$ does not create an optimal decision rule for the lognormal case, thus leading to nonconvex indifference curves in the (Q_1, Q_2) space. However, the set $Q_1 = Ex$ and $Q_2 = \sigma$ $(log\ x)$ is an optimal decision rule for the lognormal distribution. In fact, the indifference curves in the new (Q_1, Q_2) space are convex, as required.

THEOREM 1: *Let $F(x)$ and $G(x)$ be the cumulative distributions of two alternative options. Then $E_F U(x) > E_G U(x)$ for all concave utility functions U if and only if $\int_{-\infty}^{x} [G(t) - F(t)]dt \geq 0$ for all x, and for some x_0 a strong inequality holds.*[11]

THEOREM 2: *Let x be lognormally distributed. Then the necessary and sufficient conditions for $E_F U(x) > E_G U(x)$ for all concave U, are $E_F(x) > E_G(x)$, and $\sigma_F^2 \equiv variance_F [log\ x] \leq variance_G [log\ x] \equiv \sigma_G^2$.*

PROOF:

The quantiles of order p of the normal distribution (Z_p) and the lognormal distribution $(Z_{\Lambda p})$ are given by[12]

$$Z_{\Lambda p}(F) = \exp\{\mu_F + Z_p \sigma_F\}$$
$$Z_{\Lambda p}(G) = \exp\{\mu_G + Z_p \sigma_G\}$$

where Z is a normally distributed variable with mean zero and variance 1, and Z_Λ is a lognormally distributed variable. And $\mu_F = E_F \cdot (log\ x)$, $\mu_G = E_G(log\ x)$ and σ_F and σ_G are the standard deviations of *log x* under distributions F and G, respectively.

Since $E_F(x) = \exp\ \{\mu_F + (\frac{1}{2})\sigma_F^2\} > E_G(x) = \exp\ \{\mu_G + (\frac{1}{2})\sigma_G\}$ and $\sigma_F < \sigma_G$ by assumption, this implies that $\mu_F < \mu_G$. Hence there is a value Z_{p_0} for which distributions F and G intersect where $Z_{p_0} = \mu_F - \mu_G / \sigma_G - \sigma_F$. For any value $Z_p > Z_{p_0} F$ is above G and for $Z_p < Z_{p_0} F$ is below G which implies that there is only one intersection of F and G. Clearly, $\int_{-\infty}^{Z_p} [G(\mathrm{t}) - F(\mathrm{t})]d\mathrm{t} > 0$ for all $Z_p < Z_{p_0}$. However, $\int_{-\infty}^{Z_p} [G(\mathrm{t}) - F(\mathrm{t})]d\mathrm{t} > 0$ also for all $Z_p > Z_{p_0}$ since $E_F(x) > E_G(x)$ by assumption, and there is only one intersection of F and G.[13] Thus, by Theorem 1, $E_F U(x) > E_G U(x)$

---

[9] The choice among uncertain prospects depends, of course, on the information provided regarding the investor's preference and also the statistical distribution of the rates of return. According to the information provided, all the possible prospects can be divided into two sets: the efficient set and the inefficient set. One can construct the two sets by means of alternative decision rules, called efficiency criteria. Every decision rule which does not contradict the basic assumptions (or information) is a sufficient rule. However, the sufficient rule which minimizes the efficient set is a sufficient and necessary rule for a partial ordering of uncertain prospects and therefore an optimal criterion.

[10] See Paul Cootner and Feldstein.

[11] For proof see Hadar and Russell, Hanoch and the author (1969), and Rothschild and Stiglitz.

[12] See J. Aitchison and J. A. C. Brown.

[13] Note that $E_F(x) - E_G(x) = \int_{-\infty}^{\infty} [G(\mathrm{t}) - F(\mathrm{t})]d\mathrm{t} > 0$.

for all risk averters. It is easy to demonstrate that $[E(x)-\sigma\ (\log x)]$ is an optimal criterion for the lognormal distribution, i.e., the conditions of the theorem are also necessary conditions.

From the above analysis it can be seen that $[E(x)-\sigma(x)]$ is not an optimal criterion[14] for lognormal distributions of return although it is a sufficient rule. The inequality $\sigma_F^2(x)<\sigma_G^2(x)$ implies that[15]

$$(5)\quad e^{2\mu_F+\sigma^2_F}\cdot(e^{\sigma^2_F}-1)<e^{2\mu_G+\sigma^2_G}\cdot(e^{\sigma^2_G}-1)$$

Given that $E_F(x)>E_G(x)$, implies[16] that $\sigma_F^2(\log\ x)<\sigma_G^2(\log\ x)$ and therefore $E_FU(x)>E_GU(x)$ for all concave U (see Theorem 2). However, the conditions of Theorem 2 may apply even if $\sigma_F^2(x)>\sigma_G^2(x)$, and therefore the $[E(x)-\sigma^2(x)]$ criterion is a sufficient but not a necessary rule for the lognormal distributions.

Feldstein shows that indifference curves in the $[Ex-\sigma x]$ space are not convex. Tsiang maintains that this paradox stems from neglecting the skewness in the $\{Ex-\sigma x\}$ analysis. We maintain that Feldstein's paradox, as well as other possible paradoxes, stem from applying a nonoptimal decision rule. And, in fact, indifference curves in the $[Ex-(\log x)]$ space are convex, as required.

Let us assume like Feldstein that the utility function $U(x)=\log x$ and that x is lognormally distributed. Thus $\log E(x)=E(\log x)+\frac{1}{2}\sigma^2(\log x)$. Hence

$$E(\log x)=\log\ (Ex)-\tfrac{1}{2}\sigma^2(\log x)$$

Since we assume that $U(x)=\log x$, the last equation can be rewritten,

$$EU(x)=\log\ (Ex)-\tfrac{1}{2}\sigma^2(\log x)$$

One can easily check along a given indifference curve that

$$\frac{d(Ex)}{d\sigma(\log x)}=\sigma(\log x)\cdot E(x)>0$$

and

$$\frac{d^2(Ex)}{d\sigma^2(\log x)}=E(x)\cdot[1+\sigma^2(\log x)]>0$$

and the indifference curves are, in fact, convex. In other words the paradoxical result of the nonconvex indifference curves stems from the fact that a nonoptimal criterion had been used. When the optimal criterion was applied to the above problem, it was found that the indifference curves in the relevant $\{Ex, \sigma(\log x)\}$ space were convex.

IV. Conclusion

Unfortunately, the conclusions to be drawn from this paper are less encouraging than Tsiang's conclusions. In general, it is not permissible to use the *M-S* rule except under certain specific conditions. However, if the *M-S* rule is used as an approximation, it must be realized that paradoxical results similar to those obtained by Borch and Feldstein are unavoidable. Moreover, the assumptions made by Tsiang, intended to guarantee a good approximation, have been shown to be incorrect. The slope of the indifference curves in the *M-S* space is not necessarily smaller than 1. However, where a nonoptimal decision rule is being used, it may result in indifference curves of an irregular shape. The use of an optimal decision rule will, in general, guarantee convex indifference curves, as required. Tsiang's criticism of Tobin's demand for money analysis is also invalid since it relies on the incorrect assumption that the slope of the indifference curve is smaller than 1.

REFERENCES

J. Aitchison and J. A. C. Brown, *The Lognormal Distribution,* Cambridge 1963.

K. Borch, "A Note of Uncertainty and Indifference Curves," *Rev. Econ. Stud.,* Jan. 1969, *36,* 1–4.

P. H. Cootner, *The Random Character of Stock Market Prices,* Cambridge, Mass. 1964.

M. S. Feldstein, "Mean Variance Analysis in the Theory of Liquidity Preference and Portfolio Selection," *Rev. Econ. Stud.,* Jan. 1969, *36,* 5–12.

---

[14] We denote the variance of the random variable $\log x$ by $\sigma^2\ (\log x)$ or just by σ^2. However, the variance of the return x is always denoted by $\sigma^2(x)$.

[15] Aitchison and Brown.

[16] $E_F(x)>E_G(x)$ implies that $(\mu_F-\mu_G)>\frac{1}{2}(\sigma_G^2-\sigma_F^2)$. Taking *log* of both sides of (5), and using these results one finds that $\sigma_F^2(\log x)<\sigma_G^2(\log x)$ (or in short notation $\sigma_F^2<\sigma_G^2$). For further details, see the author.

J. Hadar and W. R. Russell, "Rules for Ordering Uncertain Prospects," *Amer. Econ. Rev.* Mar. 1969, *59*, 25–34.

G. Hanoch and H. Levy, "The Efficiency Analysis of Choices Involving Risk," *Rev. Econ. Stud.*, July 1969, *36*, 335–46.

——— and ———, "Efficient Portfolio Selection with Quadratic and Cubic Utility," *J. Bus., Univ. Chicago,* Apr. 1970, *43*, 181–89.

H. Levy, "Stochastic Dominance Among Log-Normal Prospects," *Int. Econ. Rev.*, forthcoming.

——— and M. Sarnat, *Investment and Portfolio Analysis,* New York 1972.

H. M. Markowitz, *Portfolio Selection: Efficient Diversification of Investment,* New York 1959.

J. P. Quirk and R. Saposnik, "Admissibility and Measurable Utility Functions," *Rev. Econ. Stud.*, Feb. 1962, *29*, 140–46.

M. Rothschild and J. E. Stiglitz, "Increasing Risk: I. A Definition," *J. Econ. Theory,* Sept. 1970, *2*, 225–43.

——— and ———, "Increasing Risk: II. Its Economic Consequences," *J. Econ. Theory,* Mar. 1971, *3*, 66–84.

P. A. Samuelson, "General Proof that Diversification Pays," *J. Finance Quant. Anal.,* Mar. 1967, *2*, 1–13.

———, "The Fundamental Approximation Theorem of Portfolio Analysis in Terms of Means, Variances and Higher Moments," *Rev. Econ. Stud.,* Oct. 1971, *38*, 537–42.

J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ Stud.*, Feb. 1958, *25*, 65–85.

———, "Comment on Borch and Feldstein," *Rev. Econ. Stud.*, Jan. 1969, *36*, 13–14.

S. C. Tsiang, "The Rationale of the Mean-Standard Deviation Analysis, Skewness Preference, and the Demand for Money," *Amer. Econ. Rev.*, June 1972, *62*, 354–71.

The Rationale of the Mean-Standard Deviation Analysis: Reply and Errata for Original Article

By S. C. TSIANG*

Reply to Karl Borch

1) Borch here essentially repeats his old point that the *E-S* indifference map is incapable of giving an unconditionally consistent representation of the preference ordering of uncertain returns. This is an incontrovertible truth which I have fully recognized in my original article.[1] He is, however, reluctant to grant my point that, nevertheless, for investors who normally undertake only small risk relatively to their total wealth the *E-S* analysis remains a highly useful approximation.

Let me take this opportunity to clarify this point by a numerical demonstration using Borch's own examples. He gives an example of three gambles A, B, and C. Gambles A and B are assumed to be equivalent (on the same indifference curve). Gambles B and C have the same mean and the same variance (2, 4). Yet C is so constructed that it is axiomatically preferrable to A and hence should be preferable to B also, even though B and C have identical means and variances. This is regarded as a proof that a consistent preference ordering of all probability distributions cannot be represented by a utility function of mean and variance, i.e., $U(E, V)$.

A casual observation would reveal why gamble C should be preferred to gamble B even though they have identical means and variances. While C is symmetric, B has a negative skewness as measured by a negative third central moment of $\bar{m}_3 = -13.6$. Since typical persons would have a preference for positive skewness, a negative skewness is therefore a disadvantage.

There are, however, very common circumstances under which an investor would hardly be concerned with the skewness at all. These circumstances involve the initial wealth of the investor and the risk he has already undertaken. The trouble with Borch's argument is that in giving these cute classroom examples he never specifies what these initial circumstances are. He seems to assume that the one or two dollar mean values of the gambles constitute the only wealth for the player (investor) concerned. This neglect causes him to miss my point entirely. Once these pieces of information are filled in, the difference between our points of view would be immediately clarified.

Let us assume for argument's sake that these gambles are to be played by a person who has initially a riskless wealth of \$10,000. Let us assume that his utility function is of the negative exponential type, i.e., $B(1-e^{-\alpha y})$, where y represents the net wealth. Furthermore, let us assume that he is reasonably ambitious (or greedy) about wealth so that a tenfold multiplication of his present wealth would not send him right into a state of utter bliss (full contentment), but would bring him no closer than, say, 1 percent away from that blissful state. Thus α must be taken as a small number of the order of magnitude of, say, roughly $1/2y_0$, where y_0 is the level of wealth that he is accustomed to.[2]

Thus game B would increase his expected utility from the initial level of $B(1-\exp(-\frac{1}{2}))$ to

$$B - B\exp\left(-\frac{10{,}002}{20{,}000}\right)\left(1+\frac{4}{8}\,\text{x}\left(\frac{1}{10}\right)^{8} - \frac{(-13.6)}{48}\,\text{x}\left(\frac{1}{10}\right)^{12} + \ldots\right)$$

---

* Professor of economics, Cornell University.

[1] See the author (1972, pp. 354–55).

[2] See the author (1972, p. 358). If he is more ambitious (or greedy) and aspires to be a multimillionaire, we must assign an even smaller value to the parameter α. Thus it can be easily deduced that the more ambitious and greedy the investor is with respect to material wealth, the less aversion for risk or preference to skewness he would have.

whereas game C would raise it to

$$B - B\exp\left(-\frac{10{,}002}{20{,}000}\right) \cdot \left(1 + \frac{4}{8} \text{ x } \left(\frac{1}{10}\right)^{8} - 0 + \dots\right)$$

Since the marginal utility of one dollar of sure wealth in this case is roughly $\alpha B e^{-\alpha y} = (1/20{,}000)B \exp(-10{,}002/20{,}000)$, therefore, the difference in monetary values of the two gambles to this particular person is only \$13.6/24 x $(1/10)^8 = 5.7$ x $(1/10)^7$ cents, which I hope Borch would agree is not a sum worth quarreling about even before the devaluation of the dollar.

Borch claims that he can easily escalate his example to demonstrate that, of two gambles with identical mean-variance-skewness triples, one can be so constructed as to be clearly superior to the other. I have no doubt about his ability to do that. But it is just as easy to escalate our numerical illustration by extending the above expansion of the expected utility function to include the fourth and higher central moments. It will be easy too to demonstrate that the differences in fourth or higher moments would be of rapidly diminishing monetary values to the person concerned as compared with the already negligible value of skewness.

2) Borch next uses an example of *gamma* distributions to demonstrate that two (E, V) vectors supposedly on the same indifference curve cannot be of equal expected utility, if the two *gamma* distributions are constructed in the way he indicated. The difficulty of handling *gamma* distributions by means of E-V analysis or E-S indifference curves is again due to the degrees of skewness of these distributions, which are closely linked with their means and variances. The problem is essentially the same as that of the previous example.

Borch, moreover, intends to use this example to demonstrate that my proposition that positive skewness is an attractive property cannot be generally valid. This demonstration, however, is based upon an obvious misunderstanding. He argues that the skewness of a *gamma* distribution decreases with an increase of the parameter n in his formula for *gamma* density. This is true if "skewness" is measured by the pure number measure $\mu_3 = \bar{m}_3/S^3$. Measured by the third moments about the means, however, it necessarily *increases* with n given α, for $\bar{m}_3 = 2(n+1)/\alpha^3$. Thus μ_3 decreases with n only because S^3 increases faster with n, as $S = \sqrt{(n+1)}/\alpha$. Therefore, $\mu_3 = \bar{m}_3/S^3 = 2/\sqrt{n+1}$, which decreases with n.

In the case of the two *gamma* densities constructed according to his specification, viz., that the two densities share the same α but one has a greater n than the other, the density with a great n must have greater mean and variance and also a greater third central moment, $\bar{m}_3$. Since the basic Taylor expansion of the expected utility function is

$$E[U(y)] = U(\bar{y}) + U''(\bar{y})\frac{S^2}{2} + U'''(\bar{y})\frac{\bar{m}_3}{3!} + \dots,$$

therefore, with U''' understood to be positive, $\bar{m}_3$ must be a desirable property. Thus the greater $\bar{m}_3$ of the *gamma* density with greater mean and variance would help to explain why it is preferable to the other, just as in the previous case of two Bernoulli distributions.

My proposition that skewness is desirable is certainly generally valid, so long as U''' is positive. Even if skewness is to be measured by $\mu_3 = \bar{m}_3/S^3$, this proposition still holds, provided other moments, especially S, remain the same.

3) Borch then reminds us that we should not build the E-V analysis on the basis of the fact that, when all stochastic variables have distributions belonging to the same two parameter family of distributions, there will be a one-to-one correspondence between the two parameters and the mean and the variance, and, hence, a preference ordering over this set of stochastic variables will be a preference ordering over the set of (E, V) vectors. For the return of a portfolio of two or more assets with stochastic returns be-

longing even to the same family of distributions will in general belong to a different type of distribution unless the distributions of the component assets all belong to the same member of so-called "stable" distributions. The only member of this class of distributions that has a finite variance is the normal distribution. Even normal distributions, when they are truncated and modified by insurance, stop-loss arrangements, progressive taxation, etc., would no longer be stable in this sense, nor normal for that matter.

This is a very pertinent reminder in view of Bierwag's and Levy's comments, which will be discussed in the following sections. But it is fully recognized in my paper. In fact this is why I emphasize that the E-V analysis, if it is to be useful for portfolio analysis, must to some extent be "distribution free."[3]

4) Of course, to be completely distribution free is inherently impossible. Since my analysis is based upon the Taylor expansion of the expected utility function, its validity depends upon the convergence of the series. I have already pointed out that for Paretian distributions with infinite moments, such as the Cauchy distribution, this analysis is of course inapplicable. Borch has now ferreted out another special family of distributions, viz.,

$$f(x) = \frac{1}{24}\left[1 - \alpha \sin 4\sqrt{x}\right] \exp\left(-4\sqrt{x}\right), \qquad (0 < \alpha < 1)$$

all the members of which have identical moment sequences, viz.,

$$m_n = \frac{(4n+3)!}{6}$$

which is independent of the only parameter α. A preference ordering based on mean, variance, skewness, etc., is then out of the question. In fact, the Taylor expansion of the expected utility function involving this density would not converge with the successive moments rapidly increasing in a factorial function as above.

However, I would not regard this as a fatal defect of the E-S analysis at all in its limited field of application as I suggested. For the E-S analysis is meant only to be a practical method of portfolio analysis for investors who regularly take rather small risks relative to their total wealth. It is not meant to be a universally valid mathematical theorem on preference ordering of all stochastic variables, or one that may be regarded as conclusively disproved by a single counterexample which mathematicians can dream up in their pipe smoke. Unless Borch can point out that there are actually investments in the financial market with returns belonging to this type of density function, we need not even be bothered with it.

Reply to G. O. Bierwag

1) Bierwag takes issue with me chiefly on my proposition that the indifference curves in the (E-S) half-plane, in so far as such analytical device is justified, must have slopes less than 45°. He contends that this proposition is quite false. To support his contention, he attempts to derive the slope of indifference curves by using an artificial example of an investor investing all his wealth in two assets, one riskless and the other risky. He lets x_0 and x_1 stand for the actual number of dollars invested in the riskless and the risky assets, respectively, without imposing any wealth constraint or sign constraint on them.

If the derivation of his equations are correct, then equations (3) and (4) combined would give the slope of indifference curves as

$$\frac{dE}{dS} = \frac{\mu\sigma}{\sigma^2 + \mu/x_1}$$

where μ and σ are the mean and the variance of the random return of the risky asset, here assumed to be *gamma* distributed. He proclaims that this proves that the E-S indifference curves can have slopes greater than 45°, since if we let x_1 approach infinity, dE/dS would approach μ/σ, which is greater or smaller than one according as $\mu \gtrless \sigma$.

The basic question here is whether the E-S indifference curves, which are supposed to represent the subjective preference ordering over combinations of E and S, should be derived from a knowledge of the investor's subjective utility function (or scale of preference) alone, or should they also incorporate information about some specific investment opportunities on the market (yet without any wealth constraint) as is implied in Bierwag's procedure. If the latter, then how are we to derive the properties of the indifference curves of an investor who is free to invest his wealth on any of the hundreds of stocks listed in the various exchanges, and government or corporate bonds and bills, plus a host of other assets, such as real estate, time deposits, antiques, cash, and what not, which are not likely to have the same type of probability distributions for their respective returns, in spite of the heroic claims of some portfolio theorists.[4] Furthermore, if we recall the Section 3) of our discussion of Borch's comment, we should know that if there are just two risky assets, the random returns of which are not distributed according to the same family of "stable" distributions (in this connection, it should be pointed out that *gamma* distributions, which Bierwag used in his example, are not stable in this sense), then we can no longer say that ". . . expected utility can be expressed as a function of only the mean and the standard deviation of the portfolio" (p. 431). That is to say, if there are two or more risky assets (unless they all have "stably" distributed returns), the theoretical basis for his analysis would disappear.

---

[3] See the author (1973).

2) Moreover, Bierwag should have noticed the glaring contradiction between his conclusion of the first section and his conclusion in the second section. It is a good thing that Bierwag has not conceded that if Prob $[\lambda<0]=0$, then the asset concerned cannot have a negative marginal expected utility. In other words, "contamination" of the portfolio on account of having too much of that asset is impossible.

In the case of the two-asset model in his Section I, both the riskless asset, cash, and the risky one, assumed to have a *gamma* distributed return, cannot have a negative return, i.e., Prob $[\lambda<0]=0$, and, hence, cannot be a source of contamination of the portfolio. If so, it is clearly unreasonable to find that a ray from the zero origin of the E-S plane cuts any E-S indifference curve twice. For this would imply that two vectors of E and S, one of which is a scalar multiple of the other, representing two portfolios, one of which is the same scalar multiple of the other, in the two assets in question, would yield the same level of expected utility. Given the convexity of the indifference curves, this would imply that starting with the bigger of the two utility equivalent portfolios, some destruction of both assets, cash and the asset with *gamma* distributed return, would increase the expected utility of the investor. This is of course absurd.

Bierwag has now accepted all this and tried to express the same fact with his proposition that the indifference curves involving such assets must have an elasticity of less than unity. That is, a ray from the zero origin can only cut the indifference curves from below, and, hence, cannot cut them twice.

However, the indifference curves he derived in Section I have a slope of zero when $x_1=0$ and a slope of μ/σ when $x_1=\infty$, i.e.,

$$\frac{dE}{dS}=\frac{\mu}{\sigma}+\frac{1}{\sigma}\left(\frac{-\mu^2}{x_1\sigma^2+\mu}\right)$$
$$=\begin{cases}0\\ \mu/\sigma\end{cases},\quad \text{when } x_1=\begin{cases}0\\ \infty\end{cases}$$

Thus any ray with a tangent between 0 and μ/σ would cut all the indifference curves it encounters at two points except the one it is tangent to. How can he reconcile this fact with his proposition that these indifference curves must have an elasticity less than unity?

I have not fully thought out an explanation for this contradiction yet. One tentative conjecture would be that the difficulty probably arises chiefly because Bierwag tries forcibly to apply the E-S analysis to obviously skewed distributions, such as the

---

[4] For example, see M. Blume, Eugene Fama, Benoit Maudelbrot.

gamma distributions, not realizing that its skewness is bound to lead to unreasonable conclusions, once the ratio of S to E becomes fairly large and the influence of skewness and other higher moments on the expected utility can no longer be safely neglected. It remains, however, for Bierwag himself to make the big effort to solve these puzzles.

Reply to Haim Levy

Levy criticizes me on three counts: 1) even if my assumptions for a good approximation are valid, one cannot apply the mean-variance analysis while ignoring all higher moments; 2) the slope of the relevant indifference curves is not in general less than 1, and, hence, my criticism of Tobin's demand for cash analysis is invalid; and 3) contrary to what Feldstein has pointed out, even in the case of lognormal distribution of returns, the indifference curves that are derived in terms of the relevant parameters are in fact convex. We shall deal with these criticisms in this order.

1) He first takes my expansion of the negative exponential utility function

$$E[U(y)] = B - Be^{-\alpha\bar{y}}\left[1 - \frac{\alpha^2S^2}{2!} - \frac{\alpha^3S^3}{3!}\mu_3 + \frac{\alpha^4S^4}{4!}\mu_4 - \dots\right]$$

and points out that even with $\alpha S = 1/2 < 1$, the third moments can exert a greater influence on the expected utility than the second. And he gives an example of an extremely skewed Bernoulli distribution to prove his point.

Now any careful reader of my article can testify that I have never asserted there that so long as $\alpha S < 1$, the expected utility can be reasonably approximated by just taking into consideration the first two moments. Rather, what I said there is that " . . . if for instance the risk under our consideration . . . ranges only from zero up to, say, 10 percent of the individual's expected value of total wealth, then, assuming $k = 1/2$ roughly, $\alpha S \leqq 1/20$. The coefficient for the pure measure of skewness . . . would be absolutely less than 1/48,000 and that for peakedness would be less than 1/3,840,000. It would seem, therefore, a fair approximation of the expected utility for most practical purposes can be obtained by considering the mean and the variance only" (1972, p. 359). I just don't understand how Levy could say that by setting $\alpha S = 1/2$, he is setting it "smaller than suggested by Tsiang."

With αS set at 1/20 for his first numerical example, the influence of skewness would surely be much smaller than that of the variance in spite of the extreme skewness of the distribution.

Of course, Levy can point out that since μ_3 is not limited in absolute value, provided we choose a sufficiently more extremely skewed distribution, we could always make the influence of skewness on expected utility greater than that of variance. This is of course true. But as I pointed out above in my reply to Borch, my suggestion that if total risk borne by an investor constitutes only a small proportion of his total wealth, the mean-variance analysis is still useful as a practical criterion for portfolio allocation is not meant to be a universally true mathematical theorem on the preference ordering of all uncertain returns. In the financial investment market, an investment return with a skewness ($\mu_3 = m_3/S^3$) of 3 is already considered very high. The example which Levy gave sounds more like a lottery ticket than any financial asset. It is rather irrelevant to "cautious portfolio investors who normally assume rather small risks relative to their total wealth." Of course, entrepreneurs who are prepared to risk a major proportion of their total wealth may indeed undertake risky projects with a distribution not unsimilar to that example given by Levy, perhaps in oil or uranium exploration. For such extremely skewed investment projects, the successive moments may increase so fast that the expansion of expected utility may take many terms to converge, or not to converge at all. In such cases, I have certainly recognized that the mean variance analysis is not adequate but must be supplemented or abandoned. (See the author, 1972, p. 368.)

Recognizing that the *E-S* analysis has a limited application to portfolio analysis for small risk takers, I proposed a rule of caution to guard against absurd results, viz., that we must never draw the *E-S* indifference curves with an upward slope of more than 45°. If Levy had adhered to this rule, he would certainly not get the paradoxical results by applying the *E-S* analysis to his second and third examples.

In this second example, he gives two rectangular distributions y_1 and y_2, of which the lower limit of the range of the second is greater than the upper limit of the first. Thus by the axiom that more wealth is always preferred to less wealth, the second distribution must be preferred to the first. Yet he so constructed the example that the second distribution has not only a greater mean (as it must) than the first, but also a greater standard deviation. Then he claims that according to the *M-S* criterion, y_1 cannot be said to be dominated by y_2." With this kind of reasoning, he proclaims that " . . . the *M-S* analysis cited is invalid" (p. 436).

A similar argument was used with another pair of rectangular distributions with the same characteristics to demonstrate once again that even under conditions favorable for its application . . . "the *M-S* analysis nevertheless leads to paradoxical results" (p. 437).

In these demonstrations, Levy has completely ignored my rule of caution that *E-S* indifference curves should never be drawn with slope greater than 45°. If he had followed this rule, he would have discovered that the paradoxical results would certainly not arise with the proper application of the *E-S* analysis. For if he plots the y_1 and y_2 of his second example in the *E-S* space, he would find that the line joining them would have a positive slope of 100/1. Similarly, the line joining y_1 and y_2 of his third example would have an upward slope of $10\cdot\sqrt{3}/1$. According to my rule, therefore, it would be impossible for any single indifference curve to pass through either pair of these *E*, *S* vectors.

Thus his examples, instead of proving the invalidity of the *E-S* analysis, merely serve to demonstrate the necessity of my rule that in so far as the application of the *E-S* analysis is warranted, the *E-S* indifference curves should never be drawn with a slope of 45° or more.

2) In spite of the fact that he obtained these paradoxical results by ignoring my rule, he promptly turned to attack this rule.

My argument for believing that *E-S* indifference curves, insofar as their use is warranted, cannot attain an upward slope of 45° is simply as follows. The application of the *E-S* analysis is warranted only when the expected utility can be reasonably approximated by its expansion involving only the terms with the first two moments of the uncertain outcome of investment, i.e.,

$$E[U(y)] \approx U(\bar{y}) + U''(\bar{y})\frac{S^2}{2}$$

From this approximate expected utility function, the slope of the indifference curves can be derived as

$$\frac{d\bar{y}}{dS} = \frac{-SU''}{U' + \dfrac{U'''S^2}{2}}$$

If $\dfrac{d\bar{y}}{dS} \geqq 1$, it would imply that

$$U'(\bar{y}) + U''(\bar{y})S + U'''(\bar{y})\frac{S^2}{2} \leqq 0$$

i.e., the expansion of $U'(\bar{y}+S)$ would be zero or negative. This implication I regard as not permissible for any portfolio investor, for it violates the basic axiom that more wealth is always preferred to less wealth.

Levy, however, insists that this implication is quite permissible. He even tries to prove his point by constructing a numerical example with a quadratic utility function. In this example, the maximum investment outcome would only carry the investor to the very brink of zero and negative marginal utility of wealth but without actually overstepping it. But the investment outcome is assumed to be so negatively skewed that the expected value of wealth is closely behind the maximum outcome. Thus although the entire range of possible outcomes would not ever stretch into the range of negative marginal utility, an addition of one standard deviation to the expected wealth would, however, make marginal utility negative. With such an artificially constructed example, he triumphantly claims that he has demonstrated that even with a quadratic utility function, the slope of indifference curves can be greater than 45°.

Such an argument surely makes economic discussion sound like some sort of mathematical numbers game! The trouble is that Levy does not seem to take much stock in the generally accepted axiom that more wealth is always preferred to less wealth. Although in the preceding section he was implicitly invoking this axiom to declare that the results obtained from the application of the E-S criterion are paradoxical, now in attacking my proposition stated above he wants to dispense with it at his own convenience. This appears to be the source of divergence between our conclusions.

His next counterexample for the demonstration of the invalidity of my proposition again displays the same tendency in reasoning. This time he takes two more investment prospects with rectangular distributions, the ranges of which overlap. The distributions are so constructed that the cumulative distributions of these prospects intersect somewhere within the overlapping part of their ranges, such that

$$\int_{-\infty}^{101} [F_2(y) - F_1(y)]dy < 0$$

and

$$\int_{-\infty}^{\infty} [F_1(y) - F_2(y)]dy < 0$$

Thus by virtue of a "theorem," which he claims to have proved himself elsewhere,[5] that $\int_{-\infty}^{x} [F_2(y)-F_1(y)]dy \geqq 0$ for every $x < \infty$ is the necessary and sufficient condition for prospect 1 to dominate prospect 2, for every utility function which is nondecreasing and concave, he concludes that there must be one concave nondecreasing utility function U_1, for which $EU_1(y_1) > EU_1(y_2)$, and another utility function U_2 for which $EU_2(y_1) < EU_2(y_2)$. Since he has constructed the example in such a way that that $(E$-$S)$ vectors of these two distributions would form a line in the $(E$-$S)$ space with a slope greater than 1, he infers that the slope of the indifference curve derived from U_2 must have slopes greater than one, for otherwise y_2 with the lower $(E$-$S)$ vector cannot be preferred to y_1 with the higher $(E$-$S)$ vector.

The validity of this tortuous argument obviously hinges upon two things: 1) that $\int_{-\infty}^{x} [F_2(y)-F_1(y)]dy \geqq 0$ for all x is indeed the *necessary* condition for F_1 to dominate F_2 for every acceptable utility function (the *sufficiency* of this condition can be readily acknowledged); and 2) the definition of acceptable utility functions. To Levy all concave nondecreasing utility functions are equally acceptable. Thus in his proof of the necessity of this condition he makes use of a utility function that is linear (hence has constant marginal utility of wealth) up to the level of y where $F_1(y)$ and $F_2(y)$ intersect, say y_0, and for $y > y_0$ the utility function becomes a horizontal line (hence has zero marginal utility).[6] To a mathematician, such a kinked utility function starting out with a constant marginal utility, which abruptly drops to zero at an arbitrary level of wealth, may be acceptable as a respectable member of the family of nondecreasing concave utility functions, but to an economist it is totally unacceptable for portfolio analysis, for it violates the axiom of nonsatiation of the demand for wealth. His proof of the necessity of the above condition for stochastic dominance is, therefore, trivial and unacceptable.

In the present context, it is obviously trivial to say that the indifference curves can indeed turn up to a vertical line if the marginal utility of wealth can be assumed to have dropped to zero at the appropriate level of wealth. In my opinion, this is the very situation we should rule out, because it violates the basic axiom for portfolio analysis.

---

[5] See Hanoch and Levy, Theorem 2, pp. 338–39.

[6] See Hanoch and Levy, p. 338.

It is true that the above theorem quoted by Levy has also been proved by Josef Hadar and W. R. Russell for a more restrictive and more acceptable class of utility functions, viz., all concave increasing utility functions.[7] The utility function they employed to prove the *necessity* of the stochastic dominance condition mentioned above is a quadratic utility function to start with, which, however, abruptly changes into a linear function, such that marginal utility is decreasing at first at a constant rate, but, at an arbitrarily specified level of wealth, it abruptly flattens out and remains constant forever. With this type of utility function, they were able to make marginal utility decrease as sharply as they wish up to that arbitrary level of wealth without fear that it might become negative.

This kind of utility function, however, is again unacceptable for our purpose. Although we could derive indifference curves from such a utility function that are as steep as we wish them to be (short of becoming vertical) up to an arbitrary expected level of wealth, beyond that level of wealth the indifference curves would flatten out, because with constant marginal utility, marginal risk aversion would vanish. Thus the indifference curves derived obviously would not be convex throughout, but would become concave after that arbitrary level of wealth. Such nonconvex indifference curves would cause a lot of difficulties in portfolio analysis.

Anyway, Hadar and Russell's utility function, as well as Levy's, certainly does not satisfy all the four essential properties of acceptable utility functions, which I listed at the beginning of my articles as desirable: viz., (a) $U'(y)>0$; (b) $U''(y)<0$; (c) $d/dy \cdot(-U''/U')\leqq 0$, which implies $U'''>U''^2/U'>0$; and (d) $d/dy(-yU''/U'')\leqq 0$. Therefore, the above theorem, which at best had been proven under a totally different framework of references, should not be brandished about as a "proof" that indifference curves under our framework of references can have slopes greater than 1.

3) Lastly, Levy suggested that indifference curves for a *log* utility function and lognormal distributions of returns need not be nonconvex as pointed out by Feldstein, if we simply substitute the standard deviation of the logarithms of returns for the standard deviation of the returns. This is indeed true. My objection to the kind of indifference curves he suggested has already been stated above with reference to the indifference curves Bierwag derived for a negative exponential utility function confronted with *gamma* distributions of returns. That is, indifference curve systems which incorporate the information about a specific type of distributions of returns into their own configuration would be of very limited use. For instance, Levy's $[E(x), \sigma(\log x)]$ indifference curves may be useful for choosing between two lognormal distributions of returns, but would be useless to determine whether any linear combination of these alternatives will not in fact be the optimum portfolio to hold for diversification purpose. For a combined portfolio of two lognormally distributed returns would no longer have a lognormal distribution. The $[E(x), \sigma(\log x)]$ indifference curve system, which incorporates the special characteristics of lognormal distributions into its configuration cannot therefore be applied to the combined portfolio.

Surely Levy does not have sufficient ground for suggesting that $[E(x), \sigma(\log x)]$ is an optimal criterion for the preference ordering of portfolios formed by combining lognormally distributed assets, or more generally portfolios constituted of assets with different distributions.

APPENDIX

ERRATA FOR "THE RATIONALE OF THE MEAN-STANDARD DEVIATION ANALYSIS, SKEWNESS PREFERENCE, AND THE DEMAND FOR MONEY"

P. 356: (a) Equation (2″) should read:

$$(2'') \qquad \left.\frac{d\bar{y}}{dS}\right| = \frac{-SU''}{U' + \frac{S^2}{2}U'''}$$

$$(E[U] = \text{constant})$$

---

[7] See their Theorem 2′, pp. 29–30.

instead of

$$\left.\frac{d\bar{y}}{dS}\right| = \frac{-SU''}{U'}$$

(b) Following this equation, the following clause should be inserted:

which, under the assumption that S is sufficiently small, can be further simplified as

$$\left.\frac{d\bar{y}}{dS}\right| = \frac{-SU''}{U'} \tag{2'''}$$

$$(E[U] = \text{constant})$$

P. 364: (a) The second equation in the left column should read:

$$\left.\frac{d\bar{y}}{dS}\right| = \frac{-SU''}{U' + \frac{S^2}{2} U'''}$$

$$(E[U] = \text{constant})$$

instead of

$$\left.\frac{d\bar{y}}{dS}\right| = \frac{-SU''}{U'}$$

(b) The third equation in the left column should read:

$$\frac{-SU''}{U' + \frac{S^2}{2} U'''} \geqq 1$$

instead of

$$\frac{-SU''}{U'} \geqq 1$$

(c) The fourth equation in the left column should read:

$$U' + SU'' + \frac{S^2}{2} U''' \leqq 0$$

instead of

$$U' + SU'' \leqq 0$$

(d) Footnote 12 (continued from p. 363). The end of the last sentence should read:

"unless the risk of the *asset* involved is greater than its expected terminal value."

P. 369: Footnote 22, third paragraph, in the equation for $\partial S/\partial x$, the numerator on the right-hand side should be:

$$W_o(x\sigma_i^2 + w_p\rho_{pi}\sigma_p\sigma_i)$$

instead of

$$W_o(x^2 + w_p\rho_{pi}\sigma_p\sigma_i)$$

REFERENCES

G. O. Bierwag "The Rationale of the Mean-Standard Deviation Analysis: Comment," *Amer. Econ. Rev.*, June 1974, *64*, 431–33.

M. Blume, "The Assessment of Portfolio Performance: An Application of Portfolio Theory," unpublished doctoral dissertation, Univ. Chicago, 1968.

K. Borch, "The Rationale of the Mean-Standard Deviation Analysis: Comment," *Amer. Econ. Rev.*, June 1974, *64*, 428–30.

E. Fama, "The Behavior of Stock Market Prices," *J. Bus., Univ. Chicago*, Jan. 1965, *38*, 34–105.

G. Hanoch and H. Levy, "The Efficiency Analysis of Choices Involving Risk," *Rev. Econ. Stud.*, July 1969, *36*, 335–46.

J. Hadar and W. R. Russell, "Rules for Ordering Uncertain Prospects," *Amer. Econ. Rev.*, Mar. 1969, *59*, 25–34.

H. Levy, "The Rationale of the Mean-Standard Deviation Analysis: Comment," *Amer. Econ. Rev.*, June 1974, *64*, 434–41.

B. Mandelbrot, "The Variation of Certain Speculative Prices," *J. Bus., Univ. Chicago*, Oct. 1963, *36*, 399–419.

S. C. Tsiang, "The Rationale of the Mean-Standard Deviation Analysis, Skewness Preference, and the Demand for Money," *Amer. Econ. Rev.*, June 1972, *62*, 354–71.

———, "Risk, Return, and Portfolio Analysis: Comment," *J. Polit. Econ.*, May/June 1973, *81*, 748–52.

Externalities as Commodities: Comment

By F. TRENERY DOLBEAR, JR.*

In a recent issue of this *Review*, Robert Meyer employed the tools of non-linear programming in an attempt to refute a conclusion of my 1967 externality article which had once before been the basis for a discussion in these pages (see E. J. Mishan, Mark Pauly, and the author (1968)). My article showed that "a per unit tax . . . which has the twin objectives of keeping [an individual] on his starting point indifference curve and attaining a Pareto optimum is not, in general, possible" (p. 98).[1] In other words, simultaneous achievement of exact compensation and Pareto optimality requires a tax scheme which is more complicated than a constant per unit tax or subsidy. At first glance Meyer's programming formulation seems to be capable of yielding constant per unit taxes or subsidies which do result in the simultaneous achievement of Pareto optimality and exact compensation. However, a second look reveals that Meyer's scheme requires something more than constant per unit tax-subsidy payments, for example, lump sum transfers.

In the standard Pigouvian treatment with compensation as a goal, individuals begin with 1) certain endowments of factors or goods, and 2) a legal system which constrains (or in some cases fails to constrain) the use of the goods. The endowments and the law result in an allocation of commodities which would obtain in the absence of trades made for the purpose of internalizing externalities. This allocation is called a "starting point" for the application of constant per unit taxes and subsidies which induce society to move to a Pareto optimum. Changes in the quantity of the externality (from the starting point) are permitted provided that the per unit tax or subsidy accompanies each unit change in the externality commodity. Exact compensation means that after the externality adjustments are made the appropriate individual(s) must be at the same indifference level experienced at the starting point.

In Meyer's programming formulation, however, there are no equations which would give rise to a starting point. He obtains exact compensation by constraining utility for the appropriate individual(s) to be equal to a predetermined [starting point] level, p. 739. And the programming model yields per unit tax-subsidy payments which would sustain a Pareto optimum at those starting point indifference level(s). But this is not equivalent to moving from a starting point allocation of commodities to a Pareto optimum allocation, using constant per unit tax-subsidy payments to insure exact compensation.

The point can perhaps be illustrated through a reconstruction of my triangular Edgeworth box externality model (see Figure 1). For those readers who have forgotten the particulars of that model, a quick reminder is in order. The triangular shape of community budget line FF reflects a constant tradeoff in production between bread and residential heat which produces a smoke by-product. For any point Q in the box, quantities of bread and heat for X are read off the vertical and horizontal axes in the usual manner. For Y, who lives by bread alone and is subjected to X's smoke, bread consumption is the vertical distance from Q to the FF community budget line. Smoke, of course, is measured from left to right on the horizontal axis.

The analysis proceeds by identifying a starting point. This point indicates the community's allocation of goods before any externalities are internalized by per unit taxes or subsidies. If Y is assumed to have a right to clean air unless compensated by X, the starting point might be E_0 in Figure 1. Exact compensation then requires that Y be left on indifference curve I_0^y, his indifference curve through E_0. Constant per unit taxes on X's

---

* Brandeis University. I am indebted to Roger Sherman for a number of helpful comments.

[1] I have, in the earlier discussion, expanded on this point and testified to my loyalty to Pigouvian-type measures, the above conclusion notwithstanding.

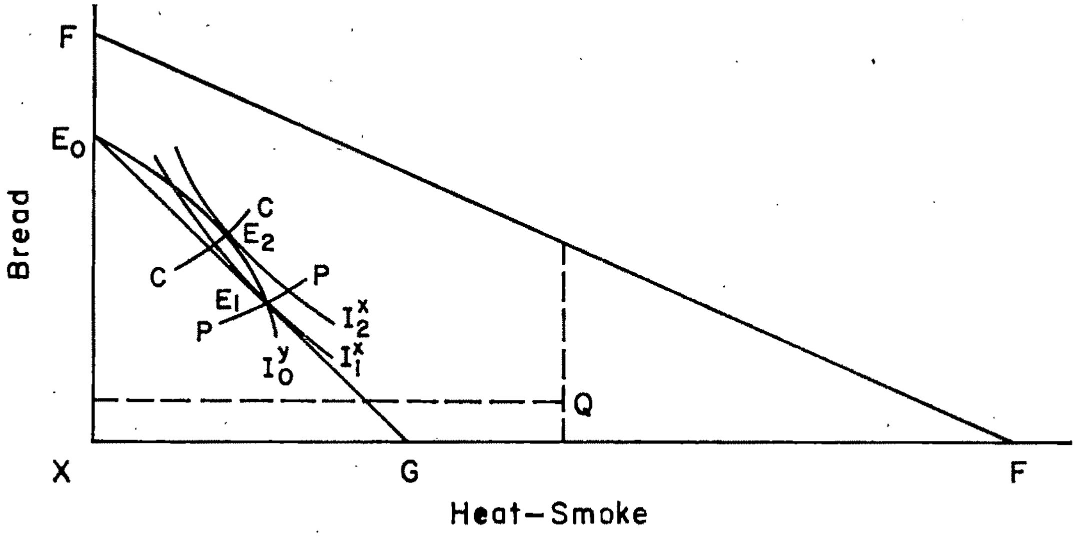


FIGURE 1

heat—to compensate for the smoke inflicted on Y—will cause X's budget line to be rotated around E_0. Higher heat taxes increase the effective price of heat and cause the budget line to be steeper. Changes in the tax rate can drive X to any point along his price-consumption curve, *pp*, which is the locus of tangencies of X's indifference curves and after-tax budget lines originating at E_0. The only exact compensation point which can be achieved through per unit taxes is E_1, at the intersection of *pp* and I^y_0. But this can easily be seen to be non-Pareto optimal since I^y_0 at its intersection with *pp* must be steeper than X's indifference curve, I^x_1, which is tangent to a ray from E_0, i.e., E_0G. In other words a constant per unit tax is unable to achieve simultaneously a Pareto optimum and exact compensation because a constant slope price line and a convex indifference curve—exhibiting a diminishing marginal rate of substitution of smoke for bread—which both start at the same point cannot be tangent, except (possibly) at the starting point.

Meyer has speculated that this result comes from " . . . some type of nonconvexity which is not stated explicitly" (p. 737). But it can easily be seen there are no nonconvexities here. Convex preferences for Y, which would lead to upside down indifference curves in a standard Edgeworth box, are in the triangular box folded in and stretched; but they are still upside down, and they will still retain the same general curvature (see my 1967 article, p. 94). Nonconvex preferences for Y would result in curves which have the same general curvature as X curves, i.e., the Y curves would no longer look upside down. Similarly, nonconvex preferences for X would result in indifference curves which did look upside down.

The impossibility result displayed above should be insensitive to mathematical technique. Meyer's translation of the problem from geometry to non-linear programming should not reverse the conclusion. A close reading of Meyer reveals that his contrary result stems from an omission of an essential ingredient from the programming formulation. His model includes a utility function, a production constraint, constraints that disallow allocations which exceed production, requirements on utility levels for individuals, and the usual nonnegativity constraints. All of that is fairly standard. (But it should be pointed out, in anticipation of the argument below, that Meyer keeps an individual at his starting point indifference level by mathematics and not by economic devices.) Meyer's innovation comes when he introduces, as his title promises, externalities as commodities in the utility functions. Each of m individuals, who consume n standard commodities, adds $n(m-1)$ externality commodities to his consumption bundle. These additions completely exhaust the bilateral possibilities for external effects between the m individuals on the n standard commodities. The quantity of one of these externality commodities is of course a function of the consumption of one of the n standard commodities by one of the $m-1$ other individuals. This completes Meyer's model.[2]

None of these constraints nor any combination of them gives rise to a starting point allocation. The only bundles of goods which appear in Meyer's formulation are the solution bundles. Thus he hasn't made use of the per unit device to trade from the starting point (determined by law) to a Pareto optimum. The failure of Meyer to specify in his programming model a starting point allocation of commodities precludes asking the very question he sets out to investigate!

Moreover, introducing the starting point into the model would upset Meyer's conclusion. His programming model yields an after-tax budget hyperplane which will not, in general, contain the starting point allocation. An additional device, such as a lump sum transfer, must be invoked first to get the individuals to the budget hyperplane. Only then will they respond to Meyer's constant per unit taxes so as to achieve a Pareto optimum with exact compensation.

The difficulty can be illuminated by reference to the triangular Edgeworth box. A Pareto optimal distribution of goods which keeps Y on his starting point indifference curve I_0^y will obtain at E_2, at the intersection of the contract curve, cc, and I_0^y. However, the after-tax budget line which will support this equilibrium, i.e., the line tangent to I_0^y and I_2^x at E_2, will not intersect the bread axis at E_0.

Meyer, and the rest of us, cannot have everything at once. Either Pareto optimality and per unit taxes mean exact compensation must be sacrificed, or Pareto optimality and exact compensation mean constant per unit taxes will not be sufficient. In short, if Meyer does not care about exact compensation, he has raised a completely empty question. But if he *does* care about compensation, his solution is a formal mathematical statement which can be seen to be wrong because it omits the starting point.

REFERENCES

F. T. Dolbear, Jr., "On the Theory of Optimum Externality," *Amer. Econ. Rev.*, Mar. 1967, *57*, 90–103.

———, "On the Theory of Optimum Externality: Reply," *Amer. Econ. Rev.*, June 1968, *58*, 529–31.

R. A. Meyer, Jr., "Externalities as Commodities," *Amer. Econ. Rev.*, Sept. 1971, *61*, 736–40.

E. J. Mishan, "On the Theory of Optimum Externality: Comment," *Amer. Econ. Rev.*, June 1968, *58*, 523–27.

M. V. Pauly, "On the Theory of Optimum Externality: Comment," *Amer. Econ. Rev.*, June 1968, *58*, 528–29.

---

[2] The reader should notice that Meyer's solution to externality problems involves transactions between *two* individuals. Thus the solution cannot be found by a decentralized market economy.

Externalities as Commodities: Comment

By TALBOT PAGE AND JOHN FEREJOHN*

In this *Review* in 1967, F. Trenery Dolbear constructed a simple general equilibrium model which demonstrated, among other things, that Pigouvian unit effluent taxes could not be expected to be both Pareto optimal and exactly compensatory to pollution sufferers.[1] Calling this result Dolbear's "negative conclusion," Robert Meyer set out in his 1971 article to generalize Dolbear's model and derive conditions under which Pigouvian taxes would achieve both goals simultaneously.

Meyer established a test which he asserted would tell us when both goals are attainable simultaneously. And Meyer suggested that when Pigouvian taxes are unable to achieve both goals simultaneously, the failure is the result of some sort of nonconvexity, p. 737. The purpose of this note is first to point out that Meyer is mistaken in thinking that the source of Dolbear's negative conclusion is in lack of convexity conditions and then to point out some of the role convexity actually plays in pollution problems.

I. Necessity of Two Instruments

One might wonder how a single policy instrument, Pigouvian per unit taxes, could be expected to achieve two disparate policy goals, except fortuitously. And indeed the reason for Dolbear's negative conclusion is the necessity for two policy instruments and not some problem with convexity. Dolbear's model is well behaved with respect to convexity conditions, as can be seen by the indifference curves and production possibility frontier of his diagram.

In Dolbear's model there are two consumers. Consumer X divides his income between bread and heat and Y spends all his income on bread. In direct proportion to X's consumption of heat, smoke is produced which Y involuntarily and distastefully consumes. We limit ourselves to the case where we start with an endowment of OB bread for X and BF bread for Y. The model is driven by X who can trade bread for heat, incidentally increasing Y's consumption of smoke (in units heat). The role of Y is completely passive: he consumes just the bread he was endowed with and the smoke blown his way.

Inside the production possibility set, which is also an Edgeworth triangle, indifference curves are drawn for X and Y.[2] Points of tangency of indifference curves form the conventional "contract" curve CC'. With initial endowment B, X will trade himself down to E_1, carrying Y to OJ_1 units of smoke consumption willy-nilly and leaving Y's bread consumption unchanged. A Pigouvian effluent tax, whose revenue goes to Y in units of compensatory bread, steepens X's effective budget constraint. Relative to initial endowment B, different Pigouvian tax rates trace out Y's "price consumption" curve PP'.

Y's indifference curve through B specifies the exact compensation requirements. Clearly, it will not in general happen that BB', PP', and CC' intersect at the same point, in which case a Pigouvian tax could be both Pareto optimal and exactly compensatory. By the geometry, the exceptional case can only happen at corner solutions or if Y's relevant indifference curve is a straight line segment.

There are two considerations in our argument that Pigouvian taxes alone cannot be expected to achieve efficiency and compensation simultaneously. The first is basically

---

* Research associate at Resources for the Future, Inc., Washington, D.C., and assistant professor of political science at California Institute of Technology, respectively. We wish to thank Roger Noll and Ralph d'Arge for helpful comments.

[1] Exact compensation for a pollutant leaves a potential pollution sufferer indifferent between the pollutant's prohibition and its allowance with compensation.

[2] The origin for X's indifference curves is 0; the origin for Y's indifference curves is F. The allocation point E_1 specifies E_1R_1 bread to Y and E_1J_1 bread to X; OJ_1 smoke to Y and OJ_1 heat to X. In the diagram, Y is indifferent between (R_1E_1 bread, OJ_1 smoke) and (R_2E_2 bread, O_2E_2 smoke).

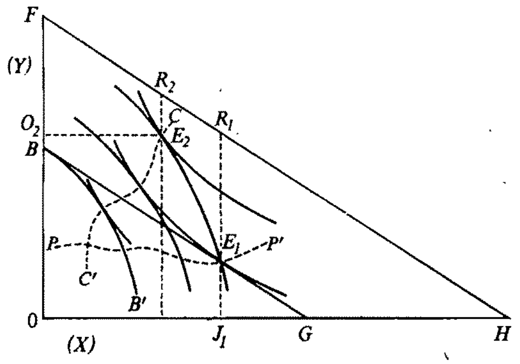


FIGURE 1

definitional. It is customary to think of a Pigouvian tax as having just a single degree of freedom, the rate itself. In Dolbear's model the set of possible Pigouvian taxes is specified by line segments hinged at E and falling between O and G. In a model of only two people, setting the tax rate for the polluter also determines the distribution of the proceeds for both participants. With a model more complicated than Dolbear's, in order to have no more than the single degree of freedom afforded by the Pigouvian tax rate, one needs to specify a fixed rule of distribution. We are not arguing that Pigouvian taxes in collection and lump sum transfers in distribution cannot achieve two policy goals, but we would count this arrangement as two policy instruments.

The second consideration is one of incentives. William Baumol has pointed out that in the case of no negotiations between polluters and pollution receivers (the "normal case") the proceeds of the effluent tax should not in general be paid to the pollution receivers in accordance with their marginal suffering. To do so would remove the incentive from pollution receivers from adopting defensive strategies which help to minimize the total social cost of pollution.[3] Consequently, even if we wanted to add to the Pigouvian tax by fine tuning its disbursement to marginal suffering, the goal of compensation would intefere with welfare efficiency. Attempting exact compensation would prevent efficiency unless payments to pollution sufferers are made conditional upon the pollution receivers taking best defensive strategies or unless there are no possible defensive strategies (the case of "pervasive pollution," which is the case in Dolbear's model). It is much simpler and more practical to use the Pigouvian tax proceeds to lower the general tax bill in a diffuse way so as to avoid perverse incentive effects in the tax distribution.

Meyer has no rule for distribution of the Pigouvian tax proceeds, and without a distribution rule his model remains incomplete as a general equilibrium model. Apparently the lack of a distribution mechanism misled Meyer in his test condition. To test whether or not per unit Pigouvian taxes can achieve exact compensation and efficiency, Meyer wrote " . . . one need only replace constraints (5) $[U^i \geq \overline{U}^i]$ by equality constraints of the form $U^i = \overline{U}^i$ . . . " (p. 739).

Meyer is maximizing a Lagrangian first with inequality constraints and then with equality constraints. It can be seen from Dolbear's model that either way this maximization leads to a definition of the same contract curve (CC' in Figure 1). With neoclassical utility functions and a negatively sloped utility frontier between X and Y, there is no need to "waste" extra utility on U^Y beyond the bare minimum $\overline{U}^Y$. Maximization of the Lagrangian will happen on the boundary $U^Y = \overline{U}^Y$ even if interior solutions are allowed. Dolbear's model passes Meyer's test and thus the test incorrectly predicts that exact compensation is possible. By omitting a distribution mechanism, Meyer seems to be confusing conditions for defining Pareto optimality (conditions independent from distribution mechanism) with solutions to a general equilibrium model (requiring tax distribution mechanisms).

There is an interesting complication. Suppose that in the presence of externalities the utility frontier bowed out sufficiently so that part of the frontier sloped upward as shown in Figure 2. Suppose also that $\overline{U}^Y$ corresponds to B, where no pollution is allowed.

---

[3] For an elaboration, see Baumol, pp. 309–12. Ronald Coase also has the example of farmers who, if they are compensated for losses due to sparks from railroad trains, have a tendency to plant too close to the tracks.

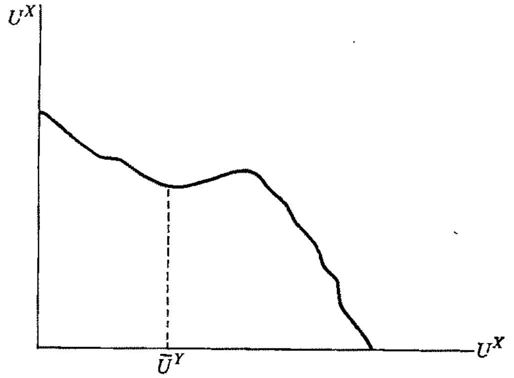


FIGURE 2

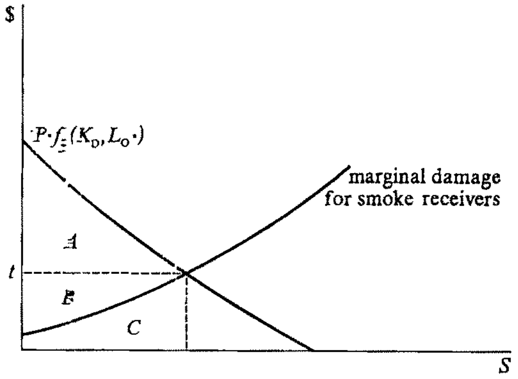


FIGURE 3

In this case the requirement of exact compensation makes welfare efficiency impossible. (Meyer's test would say that both goals are attainable because the Lagrangian max U^X subject to $U^Y \geq \bar{U}^Y$ can also be maximized on the boundary $U^Y = \bar{U}^Y$.) However, it seems that externality bads are more likely to bow the utility frontier in than out.[4]

II. Why Exact Compensation in Addition to Efficiency?

Having said this, we raise the question why a policy maker would want Pareto optimal Pigouvian taxes to have exact compensation. The point is that whether or not exact compensation suggests itself as a worthy policy goal depends in large part upon the formulation of the problem and the underlying conditions the formulation is trying to describe. In a benefit-cost analysis the question of distribution of benefits corresponds to the question of allocation of property rights in a Coasian analysis of externalities. The question of whether or not to compensate appears symmetrically. Demsetz would say the question is ethically symmetrical.

In a Pigouvian formulation, however, the situation *appears* less symmetric. In a Pigouvian analysis arising from the production sector, which is the most important case, it is natural to treat smoke as just another factor of production. For example, let quantity $Q=f(K, L, S)$, where K is capital, P the price of Q, L is labor, and S is smoke. Then the effluent tax t is analogous to the wage rate, or rental rate. For given $K=K_0$, $L=L_0$ we can draw the marginal revenue product of smoke and illustrate surpluses, as in Figure 3. Relative to a zero pollution status quo point, effluent tax t leads to a producer surplus A and consumer surplus B.

Given the practical difficulties of conditioning and distributing the proceeds $B+C$ to each individual pollution receiver on the basis of his marginal suffering, the question arises, why try?[5] For another factor of production, say L, exact compensation amounts to skimming off the labor surplus (corresponding to area B). For laborers, no one recommends exact compensation which would make them no better off working than unemployed. For laborers it is usually considered desirable not to interfere with laborers' surpluses, when they naturally appear. If we decide that a pollution receiver's surplus is not undesirable, especially since it is likely to be dispersed to consumers at large and not just to pollution sufferers, then we have no need to search for a second instrument to achieve exact compensation, nor to berate a Pigouvian tax for not doing two things at once.

---

[4] See Baumol, pp. 316–17.

[5] It seems obvious that attempting exact compensation in the typical case of several polluters and millions of receivers is attempting an administrative impossibility.

III. Convexity

In most neoclassical models, concave functions are used to insure stability and interior solutions.[6] Although Meyer has not placed any conditions on his transfer functions g^i, the behavior of the g^i turns out to be important and perhaps counterintuitive. Interestingly, in Meyer's model, *convex* transfer functions g^i assure interior solutions with Pigouvian taxes.

With $U^Y(B, S)$, Y's concave utility function between bread and smoke, and $S=g(H)$ the transfer function, linking smoke to Y's heat consumption $U^Y(B, g(H))$ may not be concave in B and H for concave function g.

A simple counterexample illustrates the point. Choose $U^Y(B,S)=B-S$ as our concave function of B and S. For $g(H)=H^{1/2}=S$,

$$U(B, g(H)) = B - H^{1/2} \tag{1}$$

is not concave in B and H, and the indifference curves bend the "wrong" way. We note that if g had been convex, Y's indifference curves between bread and heat would have been convex. The desirable convexity property of the environmental transfer function g can be stated more strongly as follows:

If $U(B, S)$ is concave, $U_2<0$ (pollution is a bad), $g(H)=S$ an increasing, convex function, then $V(B, H)=U(B, g(H))$ is concave.

A little algebra on the minors of

$$\begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

establishes the proposition.

In Figure 4(a) the function smoke $=g^1$(heat) is concave with respect to its argument heat, and the function smoke $=g^2$(heat) is convex. Figure 4(b) shows one of Y's indifference curves between smoke and bread. It is bent downward in a well behaved way. With the 45° line, Figure 4(d), this indifference curve is transformed into an indifference curve between heat and bread, depending on the functional form of g. Sufficiently concave g^2 leads to indifference curves bending up in Figure 4(c). The shear transformation between Figures 4 and 5 leaves the indifference curve convexities the same. Superimposed on Figure 6, Dolbear's Edgeworth triangle, Y^1 has the same type of convexity as indifference curves of consumer X, a situation leading to corner solutions.

What one thinks about the convexity of actual transfer functions depends partly on where one cuts off the transfer. For example, if $S=g(H)$ describes street level densities of smoke as a function of heat emissions, we may consider g nearly linear. But if $S=g(H)$ describes health damage as a function of heat emission, the function may be convex, due to diminishing returns to biological defenses.

Important policy implications follow from the degree of convexity of the environmental transfer functions. The more concave are the g, the more likely we are to recommend all or nothing policy prescriptions. Suppose, for example, water pollution damage functions are concave. Then we may want some rivers to be trout-clean and some to be industrial sewers. If the damage function is convex, it is more likely that a little pollution in all rivers is a better policy. For more convex damage functions, interior (i.e., mixed) solutions are more likely than corner solutions. Similar considerations apply to the question of whether smoking should be permitted throughout a bus or whether it should be segregated to one end.

Robert Kohn has investigated another situation which also may lead into all or nothing policy prescriptions, p. 994. Suppose the S in $f(K, L, S)$ is a vector $S=(S_1, S_2)$. The two factors of production may be sufficiently joint so that an optimal standard for both may be reached by specifying a required standard for just one. Kohn showed that this problem can be analyzed by looking at the convexity of isocost curves for abatement. It should also be noted that the convexity of air pollution damage functions partly determines whether it is wiser to limit peak episodes, by emergency measures, or to emphasize control of the chronic average pollutant levels.

---

[6] Michael Farrell has shown that the normally invoked convexity conditions are far stronger than necessary.

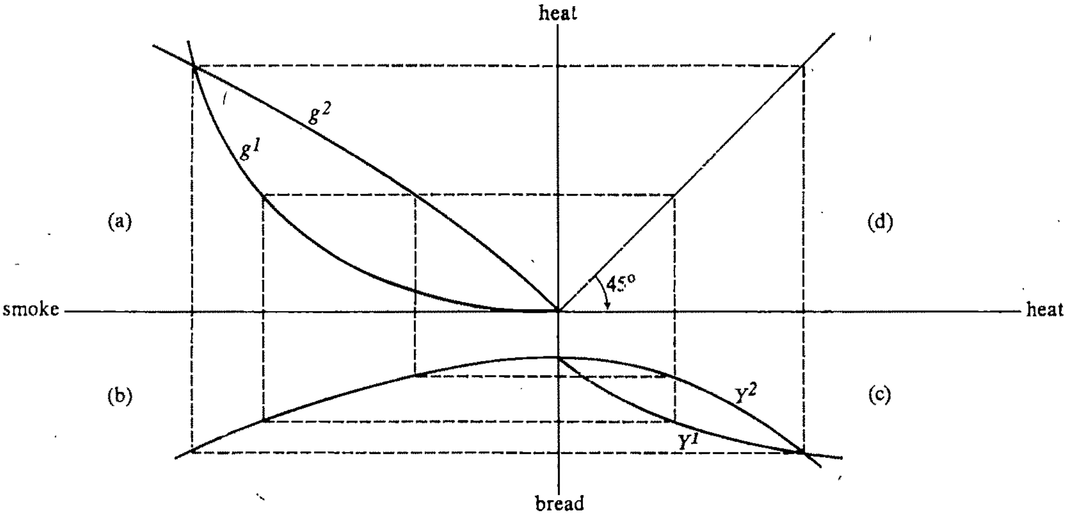


FIGURE 4

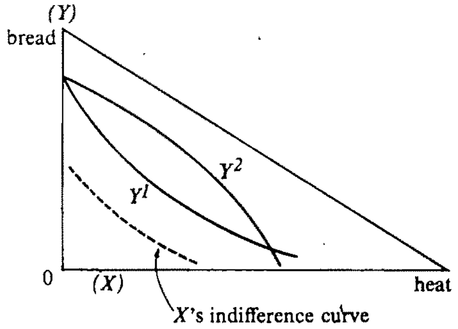


FIGURE 5

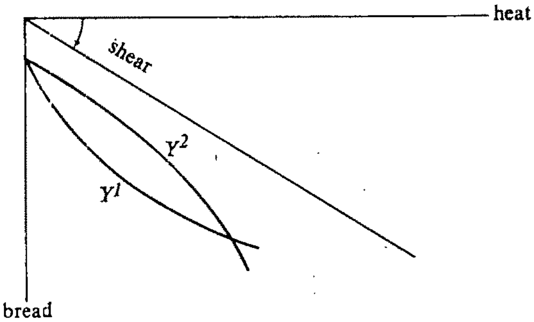


FIGURE 6

At this time there are few empirical studies which shed light on the question of convexity for g. Using cross-section data on urban areas, Lester Lave regressed total mortality rate on air pollution variables (minimum and maximum two-week averages of sulfates and particulates) and sociological variables. With both piecewise linear and quadratic specifications of the pollution variables, there was some evidence that g, here a damage function, is concave for long-run air pollution effects. The evidence is weak because the coefficients are nearly all insignificant and the sociological variable, percent over age 65, seems to be carrying the equations.

For cases where the damage function is sufficiently concave, the policy prescription will be non-Pigouvian. Depending on which corner solution is better, pollution should be either outlawed altogether or allowed without any constraint.

However, we think that in many cases g will be convex, due to diminishing returns to environmental capacity. For example, congestion is more than proportional to the number of cars on a highway.

An air pollution study provides a fragment of evidence for this second alternative. Page estimated the impact of daily levels of sulfur dioxide and particulates, in London, along with meteorological and psychological variables on a measure of perceived health, pp.

107–13. The non-linear equation

$$(2) \quad H_t = \beta_0 + \beta_1 \sum_i \lambda^i S_{t-i}^{\alpha} + \beta_2 V_t + e_t$$

was estimated where H_t is the number of people who feel worse on day t, S_t is sulfur dioxide on day t, V_t is visibility on day t, α is nonlinearity "stretching" coefficient, λ is Koyck lag coefficient, and ϵ_t is the error term. The stretching coefficient α was 1.5, indicating that this measure of perceived health is a convex function of the daily sulfur dioxide load. In another approach the same linear equation of health as a function of pollution and meteorological variables was estimated for years of high, medium, and low pollution. Decline in the pollution coefficients from the high to low years also suggested a convex damage function for pollution variables.

Most likely, transfer functions exist in a variety of forms, from concave to convex. While little has been done empirically to estimate transfer functions, from pollution sender to suffering receiver, this is an area ripe for econometric work. It is also an area ripe for theorists, for the relation of transfer functions to changes in location of receivers and senders lies at the heart of the externality problem, especially for the most interesting case of large numbers of polluters and receivers.

REFERENCES

W. J. Baumol, "On Taxation and Control of Externalities," *Amer. Econ. Rev.*, June 1972, *62*, 303–22.

F. T. Dolbear, Jr., "On the Theory of Optimum Externality," *Amer. Econ. Rev.*, Mar. 1967, *57*, 90–103.

M. J. Farrell, "The Convexity Assumption in the Theory of Competitive Markets," *J. Polit. Econ.*, Aug. 1959, *67*, 337–91.

R. E. Kohn, "Optimal Air Quality Standards," *Econometrica*, Nov. 1971, *39*, 983–95.

L. B. Lave, "Air Pollution Damage: Some Difficulties in Estimating the Value of Abatement," in A. V. Kneese and B. T. Bower, eds., *Environmental Quality Analysis: Theory and Method in the Social Sciences*, Baltimore 1972.

R. A. Meyer, Jr., "Externalities as Commodities," *Amer. Econ. Rev.*, Sept. 1971, *61*, 736–40.

R. T. Page, "Economics of Involuntary Transfers," in M. Beckmann, ed., *Lecture Notes and Mathematical Systems*, New York 1973.

Externalities as Commodities: Reply

By ROBERT A. MEYER*

Both the comments by F. Trenery Dolbear and by Talbot Page and John Ferejohn in this issue of the *Review* reemphasize an important point concerning the achievement of a Pareto optimum via constant per unit taxes. In my original paper I intended to bring out two points: 1) to set down a set of sufficient conditions under which the constant per unit taxes or subsidies envisioned in traditional discussions of Pigouvian schemes would correspond to a Pareto optimal outcome for an economy; and 2) to formally introduce "externality technologies" into a model framework (Page and Ferejohn have referred to these by what is perhaps a more apt phrase "damage functions").

As with any results developed on sufficient conditions, the assumptions about convexity are stronger than necessary to obtain the theoretical results that I stated, but they are not irrelevant if one actually contemplates calculation. All algorithms which might be employed in solving a particular applied problem seek a solution by various myopic gropings which are assured of finding a globally optimal solution only under rather stringent convexity conditions. The possibility of nonconvexity raised by Page and Ferejohn here, and earlier by Mark Pauly when he alluded to nonconstant per unit taxes or subsidies, may raise the most difficult application issues. Having the externality prices vary with quantity does not by itself raise difficulties since one could approximate non-linear pricing functions by linear segments which had constant per unit charges over each segment, but assuring that such problems do not have a multiplicity of local optima is a different matter.

In "Externalities as Commodities," I did not discuss the issue of lump sum transfers since this point was covered by E. J. Mishan, p. 527, but the reemphasis by Dolbear, his "starting point," and by Page and Ferejohn, their "transfers in distribution," is both valid and important. The only policy instrument allowed in the original Dolbear argument is the constant per unit taxes or subsidies, while the point I sought to make was a complementary one to the negative conclusion Dolbear reaches. Essentially one might be interested in asking when will a constant per unit scheme work and the answer is that one situation is the one I sketched. The "cost" of relying on such a policy to achieve both exact compensation *and* Pareto optimality is that a second policy instrument must be allowed—lump sum transfers of income. I do not feel that the solution I initially obtained is " . . . wrong because it omits the starting point," as Dolbear has asserted. My point was not meant to detract from his initial conclusion—only to complement it.

The last issue highlighted by Page and Ferejohn seems to come closest to the type of policy prescriptions that are being promulgated to cope with externalities—explicit consideration of alternative technologies (i.e., alternative damage functions). Choice among alternative externality technologies via different methods of vehicle emission control, smoke abatement, or water treatment centers attention on the economic readjustment which may be necessitated by various types of environmental "quantity" requirements. The first step is the documentation, empirically or from engineering calculations, of the levels of one or more measurable externality outputs together with the input/output levels of other factors. Once one has such quantitative data, the formulation of a least cost production problem—subject to technological constraints and given desired levels of outputs and upper (or lower) bounds on externality levels—will either indicate that there is no feasible solution or it will indicate the choice of technology which will achieve the environmental quality standards at least cost (given input prices).

---

* Associate professor, University of California, Berkeley.

REFERENCES

F. T. Dolbear, Jr., "On the Theory of Optimum Externality," *Amer. Econ. Rev.*, Mar. 1967, *57*, 90–103.

———, "On the Theory of Optimum Externality: Reply," *Amer. Econ. Rev.*, June 1968, *58*, 529–31.

———, "Externalities as Commodities: Comment," *Amer. Econ. Rev.*, June 1974, *64*, 451–53.

R. A. Meyer, "Externalities as Commodities," *Amer. Econ. Rev.*, Sept. 1971, *61*, 736–40.

E. J. Mishan, "On the Theory of Optimum Externality: Comment," *Amer. Econ. Rev.*, June 1968, *58*, 523–27.

T. Page and J. Ferejohn, "Externalities as Commodities: Comment," *Amer. Econ. Rev.*, June 1974, *64*, 454–59.

M. V. Pauly, "On the Theory of Optimum Externality: Comment," *Amer. Econ. Rev.*, June 1968, *58*, 528–29.

On Taxation and the Control of Externalities: Comment

By T. H. TIETENBERG*

In a recent issue of this *Review*, William Baumol contributes some needed insight into a turbulent and long confused area of economics—the use of specific taxes as a sufficient means of achieving an efficient allocation of resources. The article makes a rather strong recommendation for the choice of a pollution control policy. Given realistic constraints on available information, the optimal government policy would be to 1) select as a target that amount of pollution in the atmosphere which, on the basis of all available information, seems most compatible with the public interest, and 2) impose whatever tax rate on emissions necessary to yield the desired target. This emissions charge, he argues, would achieve the target pollution level at minimum cost.

This is a timely proposal because it is so compatible with the program currently being followed by the Environmental Protection Agency (*EPA*) to clean up the air in America. The *EPA* has, in essence, accomplished the first step in the Baumol proposal by establishing national air quality standards. With the exception of national controls on motor vehicle emissions and on some new stationary sources, the means by which these standards will be achieved has been left up to the states. To the state administrators on whom this responsibility has fallen, the Baumol article could well provide the rationale needed to support a massive tax system as the chief control instrument for meeting the air quality standards. Because the article could potentially have a large impact on national policy, it is important to correct an impression left by the article which at best is misleading and at worst is simply incorrect.

In this note it will be demonstrated that the alleged superiority of the Baumol proposal, while correct for the model used to analyze it, is not in general valid for actual air pollution control policy due to the lack of realism in one aspect of the model. A more general proposal, of which the Baumol proposal is a special case, will be advanced which compensates for the deficiencies of the Baumol proposal while maintaining most of its strengths.

I. The Baumol-Oates Theorem

The choice by Baumol of quality standards coupled with specific taxes to control pollution is based to a large extent upon a theorem reported by Baumol in his article but actually derived in an earlier article coauthored with Wallace Oates. The theorem states:

> Given the production of any desired vector of final outputs by the plants along the river, a tax per gallon of effluent sufficient to reduce the overall pollution content of the river to the desired level will automatically achieve this decrease at minimum total cost to all plants combined. [p. 319, fn. 15]

Although this theorem is stated in terms of water pollution, Baumol's examples, p. 318, clearly indicate his willingness to apply it to air pollution as well. The theorem implies that by using the Baumol proposal the government can achieve any desired pollution level at minimum cost by simply setting *one* per unit tax rate for each type of pollution.

This is a powerful theorem because it indicates that the implementation of the Baumol proposal will not only achieve a target level of pollution at minimum resource cost, but it will do so while incurring only minimal administrative costs. The administrator has only to set one tax rate per pollutant and this rate can be evolved by the successive application of a simple decision rule (if the observed pollution level is too high, raise the tax rate; if too low, lower the tax rate).[1]

---

* Assistant professor of economics, Williams College, and research associate, John F. Kennedy School of Government, Harvard University. I am indebted to A. Myrick Freeman III, John Sheahan, Gordon Winston, and John Todd for comments on an earlier draft.

This theorem and the policy conclusions which follow from it are correct for the model from which they were derived. Unfortunately, this model (and indeed most externality models) falls somewhat short of capturing all the essential ingredients of the process of environmental pollution. The application of the Baumol proposal in a more complicated and realistic world would not in general minimize the costs of achieving the desired pollution reduction.[2] To understand why this is so, we must examine the relationship between emission rates and air quality.

II. Emission Rates and Air Quality

The national air quality standards (the targets of the Baumol proposal) are stated in terms of concentrations in the air (for example, parts per million) rather than on a weight basis because most observed pollution damage is related to concentrations rather than weights. Emission rates, on the other hand (the source of control in the Baumol proposal), are expressed on a weight per unit time basis (for example, grams per minute). The production function which summarizes the relationship between average pollutant concentrations and average emission rates is a very complex one depending upon climatic conditions, geography and chemical reactions. Because the nature of this production process has important policy implications, a brief discussion of its essential aspects is in order.

In the meteorology literature there are a number of simulation models which describe this process in some detail.[3] Reduced to its simplest elements this process is characterized by four main phases which occur more or less simultaneously—transport, dilution, depletion, and reaction. The emissions are transported from the source by the wind with the speed and direction of the emissions flow being determined by the speed and direction of the wind. Since the wind moves in eddy currents and creates some turbulance, the emissions are scattered throughout the air mass; their concentrations are diluted.[4] Some emissions, as they are being carried by the wind, impact with the ground, buildings, etc. These emissions (lead particulates being one example) are depleted from the airshed fairly rapidly.

These transport, dilution, and depletion phases represent the meteorological aspects of the process. The remaining phase, reaction, is the most complicated and least well understood of the four. In this phase emissions react in the presence of radiation from the sun to form new substances. The two most common such reactions combine hydrocarbons and nitrogen oxides to form smog of the Los Angeles variety and the combination of moisture with a sulphur oxide (SO_3) to form airborne sulphuric acid which, among other things, tends to accelerate the deterioration of ancient monuments.

The impact of including some reasonable analytical representation of these environmental processes in a model of an economy containing pollution is that although a single per unit tax for each type of emissions will minimize the cost of reducing total emissions at the source, it will not generally minimize the cost of reducing pollutant concentrations, which are the real policy targets.[5] To minimize the costs of pollution reduction, the marginal costs of reducing the last unit of *pollutant concentration* should be the same for all emitters. This will imply that the marginal cost of *emission* reduction for all emitters be the same only if the marginal physical product (in terms of pollutant concentration) of a unit of emissions is the same for all emitters. As a point in fact, these marginal physical products are sensitive to the location of the emitter and hence will rarely be the same for all emitters.

---

[1] The adjustment process has to be somewhat more complicated than this since such a policy could theoretically lead to unbounded oscillations. As a practical matter this is not a severe limitation.

[2] Although the arguments in this comment are related to the control of air pollution, basically the same arguments can be made with respect to water pollution as well. I am indebted to Freeman for calling this to my attention.

[3] A good place to break into this literature for anyone who is interested is in the symposium proceedings edited by Arthur Stern.

[4] Mixing occurs up to some upper limit known as the mixing height. Most extreme air pollution episodes occur when this height is quite low because of the existence of a thermal inversion, a circumstance in which a layer of warm air overlays a layer of cooler air, thus limiting the upward airflow.

[5] It can also be shown that a pure tax system (i.e., one that uses only taxes and no air quality standards) which imposes the same tax rate on all emitters of a given pollutant will also be able to achieve an efficient allocation of resources only if the marginal physical products of emissions are the same for all emitters. See my 1973 paper for a formal proof.

It is not difficult to suggest why the marginal physical product of a unit of emissions will be quite different for different emitters of the same pollutant. Since emissions are transported by the wind, the location of the emitter and the direction of the wind determine where the emissions will be transported. Emitters which are downwind of the main urban air mass will make a much smaller contribution to pollutant concentrations than will upwind emitters. Hence they should pay a lower tax per unit of emissions. Forcing upwind and downwind emitters to pay the same specific tax will produce the desired concentration, but at a cost which exceeds the minimum cost means of achieving that concentration.

It is also possible, though perhaps empirically less important, that some of the other three phases are subject to intraurban spatial variation as well. Suppose, for example, in a community attempting to control smog there were two firms emitting nitrogen oxides. Suppose further that only the first of these was near a large enough source of the complementary input hydrocarbons to foster smog formation. Under these conditions a unit of emissions from the first firm would lead to a higher smog concentration than a unit of emissions from the second firm. A minimum cost reduction of smog formation implies that the first firm should incur a higher marginal cost of emission reduction than the second firm. If these two firms were to face the same tax rate (as in the Baumol proposal), their marginal costs of emission reduction would be equated and the resulting smog concentration will not be achieved at minimum cost.

III. A Zonal Price and Standards System

It becomes clear after investigating the characteristics of air diffusion processes that the marginal physical product of any emitter is determined by his location. Since this is true, unless more than one emitter occupies the same location a minimum cost price and standards policy would have to impose a different tax rate on all emitters. This would be administratively difficult at best and politically infeasible at worst. Some tradeoff between resource cost savings and administrative simplicity seems to be indicated.

The basis for fashioning a more general price and standards system is provided by the fact that the marginal physical products of a unit of emissions for two emitters emitting the same pollutant will be essentially the same if they are located in close proximity to each other. Neglecting photochemical reactions, which are beyond the modeling state of the art for the moment, existing diffusion models allow calculation of the relationship between intraurban location and the marginal physical product of a unit of emissions. It is possible, therefore, to use this information to design a zonal price and standards system which can achieve the air quality targets at very close to minimum cost while maintaining a certain degree of administrative simplicity.

The system would work as follows. The urban area would be partitioned by means of a two-dimensional grid system. Air quality standards could then be established for each zone. Through the use of existing air diffusion models (for example, see F. A. Gifford and S. R. Hanna, and D. O. Martin and J. A. Tikvart), it is then possible to standardize the tax rate paid by emitters in different locations so that a cost minimizing allocation of the control responsibility is achieved while the control authority adjusts only the numeraire tax rate.[6] This standardization procedure requires only information that can easily be derived from meteorological models; it does not require information on either control costs or damage costs.[7] All emitters of a given pollutant in the same zone would pay the same tax rate on their emissions. Emitters in different zones would pay different tax rates with those emitters in the zones where emissions contribute heavily to the concentrations in the most polluted zones would be forced to pay a relatively higher tax rate.

---

[6] The calculation of these tax rates is described in my 1974 paper.

[7] Necessary conditions for this standardization procedure to lead to an allocation which achieves the air quality standards at minimum cost are: linearity of the meteorological production function which links emissions at a source location with pollutant concentrations

The generality of this proposal comes about because an urban area can be partitioned into as many or as few zones as one would wish. The Baumol proposal represents the special case in which the urban geography is described by a single zone. More spatial detail can be introduced into the policy by using more zones to describe the same urban area. The larger the number of zones used to describe an urban area (and hence the smaller the geographic area contained within a zone) the smaller the gap between the cost of achieving the air quality standards under a zonal price and standards system (excluding costs of administration and enforcement) and the minimum cost allocation. However, it is very likely that costs of administration also increase with the number of zones. Conceptually the optimum number of zones would be that number at which the marginal saving from achieving a better resource allocation by increasing the spatial detail is equal to the marginal cost of administering that number of zones. For any larger number of zones the administrative cost would outweigh the resource allocation savings and for any smaller number the resource allocation savings from an expansion in the number of zones would exceed the additional costs of administering a larger system.

In addition to moving air pollution control policy toward a more cost effective way of accomplishing its objective, the zonal price and standards system provides one additional desirable incentive. Since urban air pollution is affected by the location of the emission sources, it is important that emitters take this into consideration when making their location choice. The Baumol proposal exercises no influence over this choice, but a zonal price and standards system provides incentives to the emitter to locate in those areas in which his emissions will cause the least harm. The environmental costs of his location are internalized.

IV. Summary and Conclusions

The Baumol proposal for achieving urban air quality standards at minimum cost is a special case of a more general price and standards system. Moving away from a uniform specific tax per pollutant toward a zonal price and standards system would lead toward a reduction in the resource cost of achieving any air quality standard. In the absence of solid empirical evidence on the cost savings derived, it nonetheless appears that the costs of administering a multizone pollution control policy do not obviously outweigh the potential savings in resource allocation. In any event, the superiority of the Baumol proposal must be demonstrated on these kinds of empirical grounds and cannot be demonstrated a priori.

REFERENCES

W. J. Baumol, "On Taxation and the Control of Externalities," *Amer. Econ. Rev.*, June 1972, *62*, 307–21.

——— **and W. E. Oates**, "The Use of Standards and Pricing for the Protection of the Environment," *Swedish J. Econ.*, Mar. 1971, *73*, 42–54.

F. A. Gifford and S. R. Hanna, "Urban Air Pollution Modeling," paper presented to the Second International Clean Air Congress of the International Union of APCA, Washington Dec. 6–11, 1970.

D. O. Martin and J. A. Tikvart, "A General

---

at a receptor location; and the underlying meteorological model must be accurate. The Gifford-Hanna and Martin-Tikvart models satisfy the linearity requirement, and recent evidence (see D. B. Turner, J. R. Zimmerman, and A. D. Busse) indicates that these models are quite accurate. For example, these models were able to explain 70 percent of the variance in the spatial distribution of the annual average concentrations of sulphur dioxide for 75 locations in the New York City area. It is worth pointing out that this high degree of accuracy was achieved in spite of the fact that the estimates are based on a rather crudely estimated emissions inventory—a source of error unrelated to the accuracy of the meteorological model. Thus even if the diffusion model were absolutely correct the explained variance would be less than 100 percent.

Atmospheric Diffusion Model for Estimating the Effects of Air Quality of One or More Sources," pap. no. 68–148, presented at the Annual Meeting of the APCA, June 1968.

A. J. Stern, *Proceedings of Symposium on Multiple-Source Urban Diffusion Models*, APC Office publ. AP-86, Washington 1970.

T. H. Tietenberg, "Derived Decision Rules for Pollution Control in a General Equilibrium Space Economy," *J. Environmental Econ. Manage.*, 1974 forthcoming.

———, "Specific Taxes and the Control of Pollution: A General Equilibrium Analysis," *Quart. J. Econ.*, Nov. 1973, *87*, 503–22.

D. B. Turner, J. R. Zimmerman, and A. D. Busse, "An Evaluation of Some Climatological Dispersion Models," paper presented to the 3d Meeting of Panel on Modeling of NATO Committee on the Challenges of Modern Society, Paris Oct. 2–4, 1972.

On Taxation and the Control of Externalities: Comment

By Earl A. Thompson and Ronald Batchelder*

In a recent article in this *Review*, William Baumol develops a simple general equilibrium model with a production externality and infers three central propositions. The first is that when the externality is of the large numbers, public goods variety and the aggregate production set is convex, a competitive equilibrium with a simple Pigouvian tax rule is a Pareto optimum. Sections I and II of this comment show that this first proposition is, in general, false. Baumol's simple model contains an unjustifiable, tacit assumption that, when dropped, leads to a new, non-Pigouvian, optimal tax rule. For a specified kind of externality, this new rule implies taxes on both recipients and imposers of the externality—a solution which Baumol emphatically rejects when there is a convex aggregate production set.

Baumol's second proposition, which is based on a reworking of a well-known example of Ronald Coase, is that a competitive equilibrium with a simple Pigouvian tax rule but nonconvexities may be Pareto nonoptimal. Section III shows that both Baumol and Coase have incorrectly applied the simple Pigouvian rule to this example. When correctly applied to the example, the simple Pigouvian tax rule (or our generalized Pigouvian rule) works; a Pareto optimal equilibrium results even though a nonconvexity is present in the example. Baumol is thus left without a proof of his general assertion that when nonconvexities are present, a competitive equilibrium with a public goods externality and a locally efficient tax system may be Pareto nonoptimal. Nevertheless, the proposition is true for a certain kind of nonconvexity, as was stated and proved by Thompson.

Finally, Baumol invents a special Pigouvian rule, which he apparently believes will solve, in principle, the problem posed by nonconvexities. Under Baumol's tax rule, there is first an a priori determination of a socially optimal allocation of resources and then a computation of marginal external losses resulting from the various outputs at this optimum. Baumol's third central inference, if we have correctly interpreted him, is that fixing tax rates at these marginal external losses—not letting them vary with the allocation of resources—will prevent the existence of inefficient competitive equilibria under nonconvexities. Section IV shows that this is not the case. Baumol's special tax rule does not, even in principle, foreclose the possibility of inefficient competitive equilibria. We demonstrate this with an example in which Baumol's rule leads to an inefficient competitive equilibrium while a simple Pigouvian rule does not.

I. Baumol's Tacit Assumption

In Baumol's model, the level of output 1 affects the cost of output 2. The producers of outputs 1 and 2 cannot contract with one another, so the effect of output 1 on output 2 is an "external effect." It is tacitly assumed that the output of each individual producer has no significant effect on the outputs of others. The choice process of Baumol's model therefore has (a) the producers of output 1 choosing their outputs for a given level of output 2, and (b) the producers of output 2 choosing their outputs for a given level of output 1. The only acceptable rationalization for this assumption is that there are many imposers and many recipients of the externality.

Baumol does assert that he is dealing with the empirically relevant, "large numbers" case, but we cannot take his brief statement to imply large numbers of *both* recipients and imposers of the externality. In many empirically relevant externalities there are only a few imposers or a few recipients, although few of both is rare for an externality of significant magnitude because the small numbers in both groups typically prevent contract costs from becoming prohibitive. The large number of costly private transactions required to internalize a significant externality when there are *either* many imposers or many recipients of the externality typically leads to the rational noninternalization of the externality (or, the simple outlawing of private transactions which internalize the externality, a government policy which may substantially reduce transaction costs). Baumol's unjustified, tacit assumption, then, amounts to the assumption that there are large numbers of *both* recipients and imposers of an externality.[1]

---

* Professor and graduate student, respectively, University of California, Los Angeles.

Note that when the externality is of the public goods variety,[2] as it is in Baumol's model and our own, it does not follow that there are necessarily large numbers of imposers and recipients. The assumption of a public good externality is logically independent of assumptions on the numbers of imposers and recipients.

In the case of one imposer and many victims of a public goods externality, the imposer is in a position to indirectly alter the outputs of the many victims by altering his own output. This cannot alter the prices relevant to the imposer since the economy is assumed to be competitive, but it may alter his taxes. In particular, under a simple Pigouvian tax rule, a single imposer may significantly affect his tax rate by inducing, with changes in his own output, changes in the outputs of the many victims. If there are many imposers and one victim, then the Pigouvian tax rule would allow a single victim to alter the outputs of the many imposers by altering his own output and thus affecting the external damage which is the basis for computing a simple Pigouvian tax rate. Pigou and Baumol did not consider these types of external tax-rate effects in specifying their simple tax-subsidy rules for eliminating externalities, and, apparently, they have been ignored in the literature on optimal taxation. Yet a model appropriate to the economic problem set up by Pigou and Baumol would allow these effects by allowing either few imposers or few recipients of an externality.

II. A Generalized Model

Baumol's necessary and sufficient conditions for Pareto optimality derived under assumptions of convexity, no corner solutions, and a competitive consumer exchange equilibrium, are

$$p_1 = c_{11} + c_{21} \tag{1}$$

$$p_2 = c_{22} \tag{2}$$

$$p_3 = c_{33} \tag{3}$$

$$p_4 = c_{44} \tag{4}$$

where $p_1, \ldots, p_4$ are market prices and marginal rates of substitution relative to leisure, $c_1(x_1)$, $c_2(x_1, x_2)$, $c_3(x_3)$, and $c_4(x_4)$ are the corresponding cost functions in terms of leisure and the second subscript represents a partial derivative with respect to the variable in the function with that subscript. In Pigouvian tradition, Baumol immediately infers, assuming convexity, that a competitive equilibrium with a Pigouvian rule will be a Pareto optimum since the price of the externality-imposing output (x_1) under a Pigouvian tax rule is equal to its private marginal cost plus the external marginal cost while the prices of the other outputs are equal to their respective private marginal costs. However, Baumol does not display any conditions for profit maximization. As a result, there is no visible determination of equilibrium outputs in his model.

---

[1] Evidence for the peculiarity of this tacit assumption is that most of Pigou's original examples involved small members of imposers or victims of his externalities. And when Pigou states the features that give rise to the problem, he says:

> One person A, in the course of rendering some service, for which payment is made, to a second person B, incidentally also renders services or disservices to other persons (not producers of like services), of such a sort that payment cannot be exacted from the benefited parties or compensation enforced on behalf of the injured parties. [ch. 9, Section 10, p. 183]

[2] The specific technology in Baumol's model generates a "public good externality" in that the amount of a particular output (smoke) which enters the production functions of others depends only upon the output (of the smoke-generating activity) of the imposer. In the case of smoke, there is no technology with which to affect the distribution of smoke among victims; that is, there is no way to effect an increase or reduction in the amount of smoke imposed on one victim without effecting the same change for all victims.

To make the profit maximizing problem explicit and thus see how outputs are chosen in a Pigouvian model, we introduce the Pigouvian profit functions:

$$\Pi_1 = p_1x_1 - c_1(x_1) - c_{21}x_1 \tag{5}$$

$$\Pi_2 = p_2x_2 - c_2(x_1, x_2) \tag{6}$$

$$\Pi_3 = p_3x_3 - c_3(x_3) \tag{7}$$

and

$$\Pi_4 = p_4x_4 - c_4(x_4) \tag{8}$$

Separate maximizations of the profit functions (5), (6), (7), and (8), imply, assuming pure competition and the existence of the appropriate derivatives,

$$\frac{\partial \Pi_1}{\partial x_1} = 0 = p_1 - c_{11} - c_{21} - x_1c_{212}\cdot\frac{dx_2}{dx_1} - x_1c_{211} \tag{9}$$

$$\frac{\partial \Pi_2}{\partial x_2} = 0 = p_2 - c_{22} - c_{21}\cdot\frac{dx_1}{dx_2} \tag{10}$$

and equations (3) and (4) above. Conditions (9) and (10) differ from Baumol's (1) and (2) in that each of the former have additional terms on the right-hand side. These terms can be set equal to zero if there are many producers of x_1 so that each of them has no significant effect on either x_2 or the marginal external damages of his industry, and if there are many producers of x_2 so that each has no significant effect on x_1.[3] Under these conditions, a Pigouvian equilibrium satisfies the Pareto optimality conditions (1)–(4), as stated by Baumol.

In general, however, either a producer of x_1 can significantly affect his own tax rate by altering the costs and thus the output of x_2, or a producer of x_2 can significantly affect x_1 by affecting his own damages and thus the tax rate faced by the producers of x_1. Consider first a situation where there is one imposer and many recipients of the externality. In this case, the individual imposer can significantly affect the costs and hence the equilibrium level of x_2 by changing his level of x_1. This brings to significance the fourth term on the right-hand side of (9), the decrease in the imposer's tax bill resulting from the reduction in output of the victims of the externality induced by a unit increase in the imposer's output.[4] Equation (9) thus violates the imposer optimality condition (equation (1)). For a Pareto optimum, the imposer must be subject to an additional tax so that any positive effect his output has on his own tax bill by reducing his tax rate disappears.

The situation where there are many imposers but only one victim of the externality brings to significance the third term on the right-hand side of (10), the reduction in the externality-loss to the victim which results from the imposers' reduction in output in response to an increase in their tax rate when the victim expands his output by one unit. It is also the increase in the Pigouvian taxes to the imposers resulting from a unit increase in the victim's output. So the conditions of equilibrium under a simple Pigouvian rule again violate the Pareto optimality conditions. For an equilibrium to be an optimum, there must be—in addition to a Pigouvian tax on the imposers' outputs—a tax on the victim's output at the marginal rate, $c_{21}(dx_1/dx_2) = -c_{21}(\partial x_1^*/\partial c_{21})c_{212}$, where x_1^* is the imposer's profit maximizing output. The two-way nature of the optimal tax is reminiscent of Coase's original two-way tax, but our result apparently holds for different reasons, as is indicated in the following section.

---

[3] The only other way to set these equal to zero is to fix the tax rate at some unalterable level. At times, Baumol so fixes the rate—at levels requiring a priori knowledge of the social optimum—but he maintains such an unrealistic tax rule only when nonconvexities appear. If Baumol had meant to maintain these fixed rates under convexity, he would never have limited his results to any large numbers case. Besides, Baumol explicitly considers the simple (i.e., variable) Pigouvian rule and asserts its optimality in a convexity case.

[4] The fifth term, x_1c_{211}, may also become nonzero, as a single imposer may significantly gain or lose from the effect of his output on total imposer taxes through its effect on marginal external damages at given outputs of the recipients. We have, however, no a priori knowledge of the sign of this term.

In summary, Baumol's conclusions fall apart when we relax the tacit assumption of both many imposers and many victims of a public goods externality. In a more general environment, an efficient tax formula would include additional taxes. In particular, Pareto optimality also requires the internalization of the tax effects on others resulting from the effect one has on tax rates through his effect on the output of others by varying the magnitude of the externality. Factories should be taxed for creating smoke to the extent that such production lowers their taxes by preventing laundries from locating nearby. And laundry production near factories should be taxed to the extent that such production lowers their costs by inducing factories to create less smoke.

III. The Sufficiency of the Simple Pigouvian Rule in Coase's Example

Coase has constructed a numerical example and alleges that it shows how simple Pigouvian tax rules can result in a nonoptimal equilibrium. Baumol agrees with Coase's allegation, argues that simple Pigouvian rules are generally inappropriate when nonconvexities are present, and suggests a special Pigouvian rule in the presence of nonconvexities, a rule which requires a priori knowledge of the social optimum. In Coase's example, an equilibrium with Baumol's special Pigouvian rule will indeed be an optimum,[5] as will an equilibrium under Coase's two-way tax. However, contrary to the assertions of both Baumol and Coase, a simple Pigouvian tax rule also achieves a Pareto optimum in Coase's example. Both Baumol and Coase have done an inadequate job of measuring the external increase in real costs.

In Coase's example, there is a factory producing smoke for its neighbors and an inelastically supplied output for the market. The decision facing the owner of the factory is whether or not to have a smoke eliminator installed at a cost of $90. Those who initially live near the factory lose $100 of real benefits if they remain near the factory but can move and avoid the harmful effects of the smoke at a cost of $40. Coase states, and Baumol agrees, that the correct simple Pigouvian tax rate is $100, a rate which induces the factory to install the smoke eliminator. This is Pareto nonoptimal because it is cheaper if the people just move away from the factory. However, the true external cost is not $100. The victims of an external diseconomy must be assumed to be able to adjust to the diseconomy in computing the damages to them; $40 is thus the true external cost to the victims of the externality. A Pigouvian tax of $40 imposed upon the owner of the factory at the initial allocation produces optimal incentives; it does not induce the factory owner to install the costly smoke eliminator so the individuals move away from the factory. (We are assuming away, following Baumol and Coase, the interaction effects discussed above in Sections I and II.)

A simple Pigouvian rule in Coase's example therefore insures that an equilibrium is an optimum. Baumol and Coase are not correct in inferring the breakdown of simple Pigouvian rules from the Coase example.

The source of the Coase-Baumol error is the failure to notice that since the individuals in the example are in a corner solution, the usual marginal equality in Pigouvian discussions, which has the cost of adjusting to an added unit of external annoyance equal to the loss from absorbing the annoyance without adjustment, does not generally hold. When this usual equality holds, the external increase in costs to others can be computed by simply measuring the loss to the victims assuming no adjustment on their part. But this usual procedure does not apply when, as in the Coase example, the cost of adjusting to an external annoyance is below the cost of absorbing the annoyance without adjustment.

While there is no breakdown of a simple Pigouvian rule in the many-imposer, many-recipient case under the location-transport-cost nonconvexity discussed by Baumol, there is a breakdown under some nonconvexity. This was shown in Thompson's 1968 paper. An example occurs when the total external diseconomy resulting from the total output of a competitive industry rises very rapidly with the output at low levels of industry output and then rises only gradually, if at all, with output at moderate and large levels of industry output. Any social optimum may imply a moderate or large industry output, but there is an equilibrium at a low level of industry output (possibly zero) because the Pigouvian tax rate at that output is very high (possibly prohibitive).

---

[5] However, Baumol's apparent inference that his rule works for all cases involving nonconvexities is false, as is shown in the next section.

IV. The Insufficiency of Baumol's Rule

Consider an industry with a very large external *economy* for small outputs but no external economy for large outputs. Say the first firm entering an industry provides a road which will accommodate a few other firms, the second firm pays for telephone lines which will accommodate others at marginal costs, the third firm provides a pipeline to a water source which can be cheaply tapped by other firms, etc., until the industry grows to a size such that it pays the firms providing such services to its coproducers to collect overhead costs from them. Suppose Pareto optimality dictates a large output. Baumol's special Pigouvian tax rate is therefore zero. But, with a sufficiently elastic demand curve for the output of the industry, there is an equilibrium at a zero output since *private* costs there exceed the demand price for the output. Baumol's special Pigouvian tax rule will permit a nonoptimal equilibrium. It is interesting to note that a simple Pigouvian rule insures that an equilibrium is an optimum in this example. But, as we have already pointed out, there is no known tax rule that survives all forms of nonconvexity.

REFERENCES

W. J. Baumol, "On Taxation and the Control of Externalities," *Amer. Econ. Rev.*, June 1972, *62*, 307–22.

R. H. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, *3*, 1–44.

A. C. Pigou, *The Economics of Welfare*, 4th ed., London 1932.

E. A. Thompson, "The Perfectly Competitive Production of Collective Goods," *Rev. Econ. Statist.*, Feb. 1968, *50*, 1–12.

On Taxation and the Control of Externalities: Reply

By William J. Baumol*

I have no desire to take issue with either of the preceding comments, both of which clearly contribute significant insights on the subject. The Tietenberg comment is entirely apropos, arguing that when the damage caused by emissions varies geographically, a uniform tax rate will not minimize the social costs of achieving a preselected level of environmental quality. I certainly am guilty of some oversimplification on this matter.

One of Earl Thompson and Ronald Batchelder's basic points is that where there is a small number of emitters of pollutants in a neighborhood, a case that occurs frequently in practice, the Pigouvian solution may require some modification. For the emitting firm will then have a motivation to adjust its output of pollutants in a way that makes the tax rate more favorable, just as a monopolist benefits by adjusting his output to secure a more profitable price.

I certainly cannot disagree with this conclusion. It is a bit misleading in one respect, however. In Pigou's writings, and in much of the subsequent literature, it seems quite clear to me that the assumption of pure competition is taken to mean that both prices and tax rates are beyond the influence of the individual firm. Strictly speaking, of course, a firm selling its product in a national market that is perfectly competitive may be the only polluter in its neighborhood. But surely all of us would readily have admitted that competitive solutions, including Pigou's externalities prescription, break down when the firm can influence any of its prices, including the tax it pays on its emissions. I prefer, therefore, to take the Thompson-Batchelder discussion of this point not as a criticism of received doctrine—which is the way they apparently want to view it—but as an illuminating examination of the case in which the firm *can* influence the tax rate.

Their second major point is that the Pigouvian solution may work better than I suggest in cases where there are several local maxima. They argue that the appropriate measure of marginal social damage and thus, of Pigouvian tax at some local maximum point, is not the cost an additional puff of smoke would then impose (*ceteris paribus*) on nearby laundries, but the opportunity loss as against what it would do if the economy were really at the global optimum.[1] For in that case at an inferior local optimum the polluters would still find an inducement in the tax to move toward the global optimum. Perhaps their definition of social damage as opportunity cost will be judged to be stretching the Pigouvian concept. Certainly it is not what emerges from the Kuhn-Tucker conditions whose values are strictly local, and whose shadow prices will lead only to the local maximum.

Finally, one protest. Thompson and Batchelder attribute to me a "special Pigouvian rule" for the multiple maximum case, but I assure the reader that I do not now nor have I ever advocated any such measure. For I certainly agree with them that there is *no* known tax rule that survives all forms of nonconvexity.

---

* Princeton and New York Universities.

[1] This interpretation may be a misreading of Thompson and Batchelder. They may simply be chiding me for going along with a peculiarity of the Coase example which I had noted in my article—the premise that the laundries will stay near the source of smoke and absorb a $100 cost even though $40 will move them to a place where the air is pure. I went along with it because without this assumption Coase's example falls apart, for no intervention is needed to achieve the optimum then. Nor can the example be saved by making the cost of moving greater than the cost of staying and absorbing the smoke damage for then the Pigouvian tax which suppresses the smoke clearly yields the right solution. Perhaps it *would* have been better to leave the discussion of the Coase example at that.

Consumer Preferences and Gains from Trade

By YEONG-HER YEH*

It has been shown that free trade is superior to no trade.[1] The consumption possibility curve under autarchy would lie inside the consumption possibility curve under free trade except touching it at one point. Therefore, the situation utility possibility curve (sometimes called grand utility possibility curve) under autarchy would lie inside the situation utility possibility curve under free trade except touching it at one point. The purpose of this paper is twofold. In Section I, it is demonstrated that this relationship does not hold under the assumption that consumer preferences are identical for each individual and the indifference curves which reflect consumer preferences are homothetic.[2] Under this assumption, the situation utility possibility curve under autarchy would lie either strictly inside the situation utility possibility curve under trade or would coincide with it entirely. In Section II, it will be shown that the situation utility possibility curve under autarchy still may lie entirely inside the situation utility possibility curve under trade after we remove the assumption that individual preferences are identical and the indifference curves are homothetic.

I

In the discussion below, we will investigate the case where a country enjoys monopoly power in trade. But the method of analysis can be applied to the case where a country has no monopoly power and faces a given international price line. The same conclusion can be obtained.

In Figure 1, FF' is the production possibility curve or the consumption possibility curve under autarchy, and QQ' is the Baldwin envelope.[3] This envelope is the trade possibility curve or the consumption possibility curve under trade.

Suppose that point J on FF' is a point at which the slope of FF' is equal to the common slope of indifference curves at any tangent point on the contract curve OJ. (If individuals A and B have identical and homothetic indifference curves, the contract curve would be a straight line and the common slope of indifference curves at any tangent point would be equal to each other.) The point utility possibility curve derived from the contract curve OJ would become the situation utility possibility curve under autarchy. Any point utility possibility curve, which is derived from the contract curve of the Edgeworth box associated with any other point on FF', would lie below the situation utility possibility curve derived from OJ. For example, pick up any point $\bar{J}$ on FF'. In the Edgeworth box $\bar{J}K_2OK_1$, those indifference curves of individual A which fall in the area of $OK_4\bar{J}K_2$ will have a slope at any point steeper than that of FF' at J. Draw a line from $\bar{J}$ parallel to OJ and to intersect the vertical axis at K_3. Those indifference curves of individual B which fall in the area of $K_3OK_1\bar{J}$, will also have a slope at any point steeper than that of FF' at J. (The origin for B's indifference map is shifted from J to $\bar{J}$; B's indifference curve U_B is now represented by $\bar{U}_B$. The slope of any $\bar{U}_B$ at its intersection point with $\bar{J}K_3$ is exactly the same as that of any U_B at its intersection point with OJ.) Therefore, the common slope of indifference curves at any tangent point on the contract curve $O\bar{J}$ will be greater than the slope of FF' at J which is in turn greater than the slope of FF' at $\bar{J}$. Thus, the point utility possibility curve derived from $O\bar{J}$ would not touch the situation utility possibility curve derived from

---

* Associate professor of economics, University of Hawaii. I am indebted to Robert Baldwin and George Borts for their valuable comments. However, I am responsible for any error in the paper

[1] See Robert Baldwin, Jagdish Bhagwati, Murray Kemp, and Paul Samuelson.

[2] This assumption is usually made in the Heckscher-Ohlin trade model. See my 1968 article.

[3] See Baldwin, pp. 95–97.

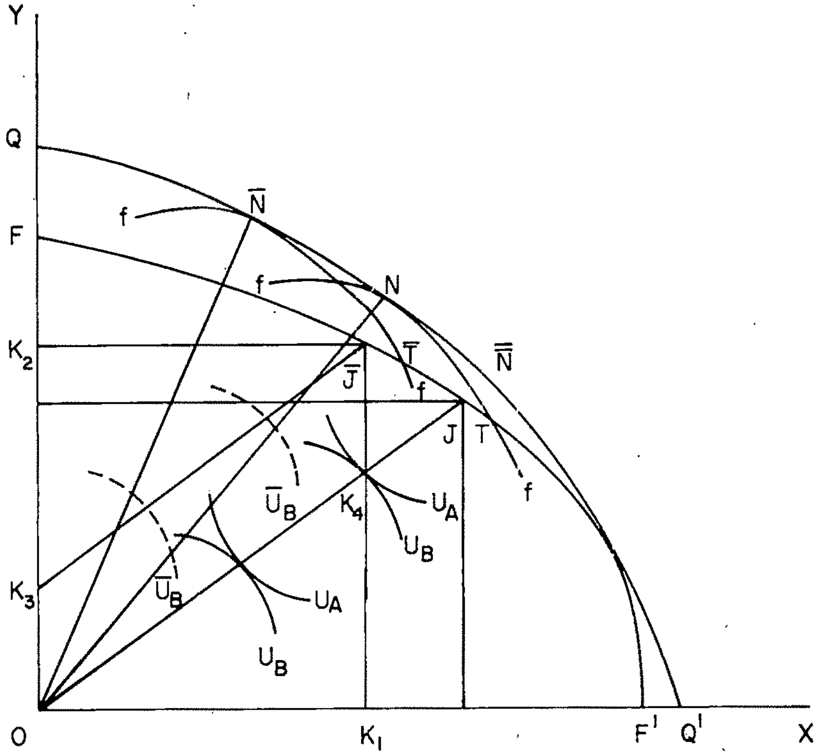


FIGURE 1. CONSUMER PREFERENCES ARE IDENTICAL FOR EACH INDIVIDUAL, AND CONSUMER INDIFFERENCE CURVES ARE HOMOTHETIC

OJ, but would lie below it. (The condition for being on the situation utility possibility curve under autarchy is that the marginal rate of substitution between goods X and Y for both A and B is equal to the marginal rate of transformation of these two goods in production.)

Next, suppose that at point N on the Baldwin envelope QQ' the common slope of indifference curves at any tangent point on the contract curve ON is equal to the marginal rate of transformation in production at T. (The latter is in turn equal to the marginal rate of transformation in trade. This is so because the slope of foreign offer curve ff at N is equal to that of FF' at T.) The point utility possibility curve derived from the contract curve ON will become the situation utility possibility curve under trade. Any point utility possibility curve which is derived from the contract curve of the Edgeworth box associated with any other point on QQ' would lie below this situation utility possibility curve derived from ON. For example, pick up any point $\overline{N}$ on QQ'. Using the same analysis as above, it can be shown that the common slope of indifference curves at any tangent point on the contract curve $O\overline{N}$ will not be equal to the slope of FF' at $\overline{T}$ (which happens to equal to the slope of the foreign offer curve ff at $\overline{N}$). In other words, the marginal rate of substitution in consumption is not equal to the marginal rate of transformation in production and the marginal rate of transformation in trade. The condition for being on the situation utility possibility curve under trade is that these three rates have to be equal to each other. Therefore, the point utility possibility curve derived from $O\overline{N}$ will lie below the situation utility possibility curve derived from ON.

The situation utility possibility curve under autarchy would lie strictly inside the situation utility possibility curve under trade. This is so because the former would lie uniformly inside the point utility possibility curve derived from $O\bar{\bar{N}}$ ($\bar{\bar{N}}$ is just any point on QQ' which contains more of both goods than point J), and the point utility possibility curve derived from $O\bar{N}$ would lie below the situation utility possibility curve under trade.

In Figure 2, R is the point where the Baldwin envelope and the production possibility curve are tangent to each other. If the consumer preferences are such that the common slope of indifference curves at any tangent point on the contract curve OR is equal to the marginal rate of transformation at R, the point utility possibility curve derived from OR will become the situation utility possibility curve both under autarchy and under trade. Therefore, there will be no gains from trade. Using the same analysis as above, we can prove that any point utility possibility curve which is derived from the contract curve of the Edgeworth box associated with any other point on FF' would lie below the point utility possibility curve derived from OR. Likewise, we also can prove that any point utility possibility curve which is derived from the contract curve of the Edgeworth box associated with any other point on QQ' would lie below the point utility possibility curve derived from OR. For example, pick up any point S on QQ'. The common slope of indifference curves at any tangent point on the contract curve OS will not be equal to the slope of FF' at U, which equals the slope of the foreign offer curve ff at S.

In summary, under the assumption that consumers have identical preferences and the indifference curves are homothetic, the situation utility possibility curve under autarchy would either lie strictly inside the situation utility possibility curve under trade, or would coincide with it entirely. In the latter case, there will be no gains from trade.

II

Let us now remove the assumption that

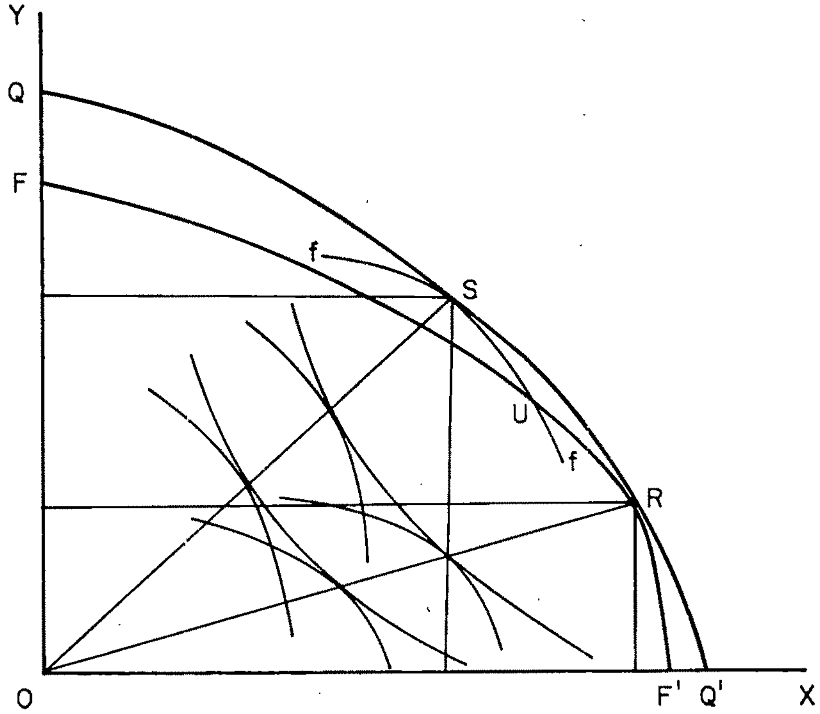


FIGURE 2.

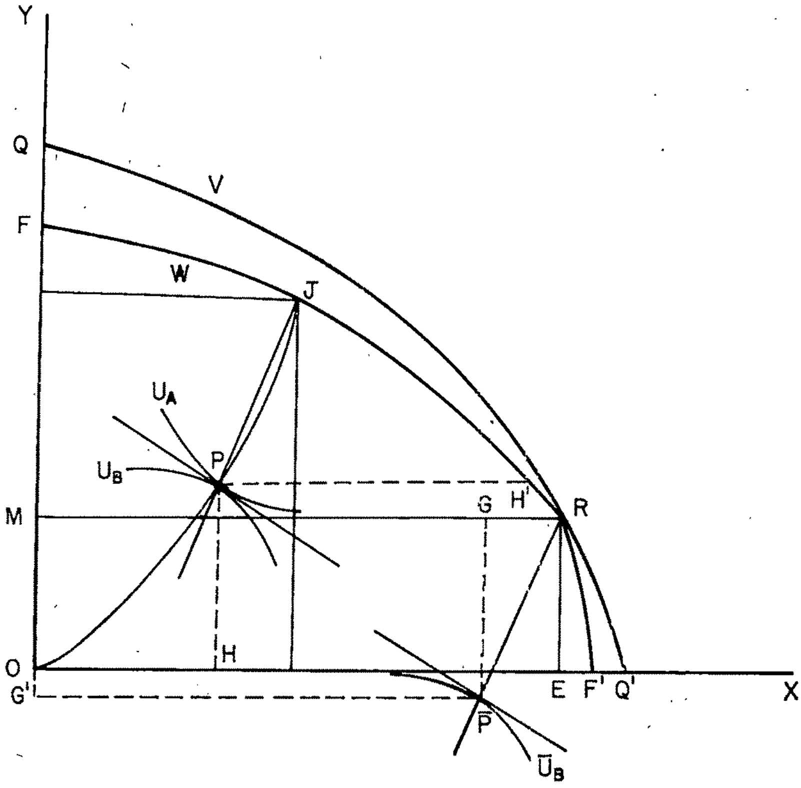


FIGURE 3.

individual preferences are identical and the indifference curves are homothetic. Instead, we just assume that neither good X nor good Y is an inferior good to the consumers. It is still possible for the situation utility possibility curve under autarchy to lie entirely inside the situation utility possibility curve under trade. This can be demonstrated as follows: In Figure 3, pick up any point J on FF'. Let us assume that we have a point P on the contract curve OJ at which the common slope of indifference curves is the same as the slope of the production possibility curve FF' at J. Thus, the point utility possibility curve derived from this contract curve OJ will be tangent to the situation utility possibility curve under autarchy. However, this does not mean that every point utility possibility curve derived from the contract curve of the Edgeworth box associated with any other output point on FF' would be tangent to the situation utility possibility curve under autarchy.

For example, the point utility possibility curve derived from the contract curve OR may not be tangent to the situation utility possibility curve under autarchy. In Figure 3, those indifference curves of individual A which fall in the area of $PHF'H'$ would have a slope at any point less steep than that of the indifference curve U_A at P. This is so because we assume that neither good X nor good Y is an inferior good.[4] By the same reasoning, those indifference curves of B which fall in the area of $\bar{P}GMG'$ would also have a slope at any point less steep than that of the indifference curve $\overline{U}_B$ at $\overline{P}$ or the indifference curve U_B at P. (Point P on U_B is shifted southeastward to point $\overline{P}$ on $\overline{U}_B$ because the origin for B's indifference map is shifted from J to R.)

---

[4] If we move directly southward from P to H through A's indifference curves, the lower and lower indifference curves which we shall cut successively will have less and less steep slopes. Similarly, A's successive indifference curves, which we shall cut as we move due east from P to H', will be less and less steeply sloped.

Therefore, the common slope of indifference curves at any tangent point in the Edgeworth box $OMRE$ will always be less steep than the slope of U_A at P. Since the slope of FF' at R is steeper than that of FF' at J and the latter in turn is equal to that of U_A at P, it is impossible to have such a tangent point on the contract curve OR that the common slope of indifference curves at this point is equal to the slope of FF' at R. This means that the point utility possibility curve derived from the contract curve OR will never be tangent to the situation utility possibility curve under autarchy.

The consumption possibility curve under autarchy, FF', lies inside the consumption possibility curve under trade, QQ', except touching it at R. For any output point (except R) on FF', for example, W, there will always be a point on QQ', such as V, which contains more of both commodities. The point utility possibility curve derived from point W would lie below the point utility possibility curve derived from point V. The situation utility possibility curve under autarchy is the envelope of the point utility possibility curves derived from each point of FF', whereas the situation utility possibility curve under trade is the envelope of the point utility possibility curves derived from each point on QQ'. Therefore, it is generally expected that the situation utility possibility curve under autarchy would lie inside the situation utility possibility curve under trade except touching it at one point, which corresponds to the nontrading point R. However, as demonstrated above, the point utility possibility curve derived from point R may not touch the situation utility possibility curve under autarchy. In this case, the situation utility possibility curve under autarchy would lie entirely inside the situation utility possibility curve under trade.

REFERENCES

R. E. Baldwin, "The New Welfare Economics and Gains in International Trade," *Quart. J. Econ.*, Feb. 1952, *66*, 91–101.

J. N. Bhagwati, "The Gains from Trade Once Again," *Oxford Econ. Pap.*, July 1968, *20*, 137–48.

M. C. Kemp, "The Gains from International Trade," *Econ. J.*, Dec. 1962, *72*, 803–19.

P. A. Samuelson, "The Gains from International Trade Once Again," *Econ. J.*, Dec. 1962, *72*, 820–29.

Y. H. Yeh, "A Note on Consumer Preferences in the Heckscher-Ohlin Trade Model," *Econ. Rec.*, Dec. 1968, *44*, 516–19.

A Note on the Price Adjustment Mechanism for Rental Housing

By LAWRENCE B. SMITH*

Until recently, interest in the price adjustment mechanism in the housing market has been relatively low, as most economists have envisioned a mechanism similar to one described by David Blank and Louis Winnick in the early 1950's. One of the features of this analysis is that vacancy rates are pivotal in explaining the price of housing services (rent) since rents vary inversely with the vacancy rate within a critical zone of occupancy.[1] Recently, however, doubt has been cast upon this traditional view by Frank de Leeuw and Nkanta Ekanem (1971) who found that the vacancy variable was insignificant in explaining the rent on rental dwellings after taking into account income, price and cost variables. Consistent with this finding is the analysis by Ray Fair who argues that disequilibrium is common in the housing market because the price of housing services does not adjust quickly enough to clear the market. It is the purpose of this note to support the traditional view by presenting empirical evidence on the price adjustment mechanism.

I. The Model

In the model presented by de Leeuw and Ekanem (1971), rent per unit of housing service R is a linear function of the price of capital inputs C, the price of operating inputs O, the number of households H, real income per household Y, the price level P, and the vacancy rate V, with all the variables in logarithmic form. When this model was estimated, the vacancy rate V and number of households H were insignificant or wrong-signed and hence deleted. The price of capital inputs and operating inputs variables also often performed less satisfactorily than expected. Although the authors on theoretical grounds did not anticipate the insignificance or poor performance of these variables, the insignificance of the vacancy variable is not surprising since the rate of change of rents is affected by the vacancy rate rather than the rate of change of the vacancy rate, and hence the vacancy variable is likely to be redundant in the model specified.

The price adjustment mechanism in the rental housing market can be viewed as operating in a typical stock-flow manner. At any given time there is a stock of rental housing units, providing housing services, and a demand for these services. The supply of housing services in a period is identically equal to the supply in the previous period plus the additional services provided by newly converted or completed rental units during the period, less the services lost by removals and demolitions of rental units and by depreciation of the previous stock. The demand for rental services depends primarily upon a variety of demographic variables (the number of families, age composition of the population, tendency to form nonfamily households, etc.), permanent real disposable income, the rent on rental accommodations, the cost of owner-occupied housing, the price of alternate goods and services, the cost and availability of mortgage credit associated with homeownership, and consumer preferences. These demand and supply functions interact to determine simultaneously the level of rents and stock of vacant rental units or the vacancy rate.

However, the level of rents and the vacancy rate so determined often deviate from their equilibrium or perfectly competitive levels for considerable periods of time because numerous frictions cause the housing market to adjust slowly.[2] Suppose the

---

* Professor of economics, department of political economy, University of Toronto.

[1] This analysis often assumes that the responsiveness of rents to vacancy rates declines as vacancy rates increase, and that there is some lower limit to rents beyond which the dwelling is abandoned.

[2] See Anthony Downs for a discussion of a number of causes of friction in housing markets.

initial solution is one in which rents are depressed below their equilibrium and the vacancy rate is correspondingly below its equilibrium. The excess demand reflected in the abnormally low vacancy rate will exert upward pressure on rents, and the level of rents and the vacancy rate will gradually approach their equilibrium levels.[3] Analogously, if the rent level and vacancy rate are initially excessive, the abnormally high vacancy rate will exert downward pressure on rents. Consequently, the rate of change of rents depends upon the vacancy rate. Variations in the arguments in the demand function and components of the supply equation will be reflected initially in the vacancy rate, although they may also appear to have some direct effect on the rate of change of rents if a sufficiently long time period is considered. Finally, variations in the rent and vacancy rate affect the expected profitability of new construction and hence the supply of future rental housing services, and thereby future vacancy rates and rent levels.[4]

This discussion can be simplified and tested with respect to the rent adjustment mechanism as follows. Assume that the demand for housing service D is a function of R, Y, P, and H,

$$D = f(R, Y, P, H) \tag{1}$$

and the supply of housing services S is fixed in the short run.[5] The vacancy level VL is then a measure of the excess supply or demand[6]

$$VL = S - f(R, Y, P, H) \tag{2}$$

Dividing by S gives the vacancy rate V

$$V \equiv \frac{VL}{S} = 1 - \frac{1}{S} f(R, Y, P, H) \tag{3}$$

It can now be seen that variations in R, Y, P, and H directly affect the vacancy rate. If we now hypothesize a price adjustment mechanism in which excess demand or supply determines rent variations, the rate of change of rents $\dot{R}$ is a function only of the vacancy rate, and introducing Y, P, and H into this equation is redundant, i.e.,

$$\dot{R} = g(V) \tag{4}$$

This specification may be too simple since cost push elements also directly affect rent inflation, as landlords try to pass operating cost increases on to the tenant. To reflect this an additional variable, the rate of change of property taxes $\dot{T}$ was introduced, since property taxes are the major operating expense, accounting for over 50 percent of nonmortgage operating costs.[7]

$$\dot{R} = r(V, \dot{T}) \tag{5}$$

II. The Empirical Results

Equations (4) and (5) were estimated for each of five Canadian cities, Halifax, Montreal, Toronto, Winnipeg, and Vancouver, using annual data 1961 (or as soon thereafter as data was available) to 1971, and the results are presented in Table 1. In addition, the data were pooled and combined cross-section time-series regressions for each model were estimated with city dummy variables D_1, D_2, D_3, and D_4 introduced for Halifax, Montreal, Toronto, and Winnipeg, respectively, and the results are presented in Table 2. Because of the likelihood of a response lag, a distributed lag relationship was

---

[3] See de Leeuw and Ekanem (1973) for a discussion of lags in the adjustment of rents to a variety of shocks, and Sherman Maisel for an elaboration of the role of vacancies in the housing market.

[4] For a fuller discussion and empirical estimates of some of these interrelationships see the author, and de Leeuw and Ekanem (1973).

[5] This assumption is not necessary for the argument, but it simplifies the notation.

[6] Since it is usually assumed that some vacancies are required for a smooth functioning of the market, excess supply is really the amount of vacancies in excess of the natural level, and excess demand is the deficiency of vacancies below the natural level. If we assume that the natural rate is some constant vacancy rate, variations in the vacancy rate reflect variations in excess supply or demand.

[7] Since mortgage terms are usually arranged prior to construction, financing costs can be considered to affect rents through the volume of new construction. Variations in current mortgage costs usually do not affect the operating costs of rental dwellings unless the structure is being refinanced. The other major operating costs should be included but poor data precluded this.

Table 1—Regression Results by City, Rate of Change of Rents $\dot{R}$, Annual Observations[a]

City	Constant	V	V_{-1}	$\dot{T}$	$\dot{T}_{-1}$	R^2	DW	Estimation Period
Halifax:	5.22	−.630	−.315			.69	1.82	1964–71
	(6.02)	(3.65)	(3.65)					
	2.45	−.221	−.110	.141	.071	.85	1.92	1964–71
	(1.76)	(1.01)	(1.01)	(2.35)	(2.35)			
Montreal:	3.04	−.178	−.089			.14	.54	1961–71
	(2.27)	(1.22)	(1.22)					
	1.73	−.111	−.053	.150	.075	.64	1.78	1961–71
	(1.72)	(1.09)	(1.09)	(3.33)	(3.33)			
Toronto:	4.95	−.582	−.291			.64	.69	1961–71
	(5.98)	(3.93)	(3.93)					
	4.69	−.565	−.282	.021	.011	.63	.66	1961–71
	(2.46)	(2.92)	(2.92)	(.15)	(.15)			
Winnipeg:	5.15	−.701	−.351			.83	1.36	1964–71
	(7.71)	(5.48)	(5.48)					
	4.87	−.696	−.348	.037	.019	.87	1.60	1964–71
	(7.04)	(5.61)	(5.61)	(1.18)	(1.18)			
Vancouver:	5.70	−.732	−.366			.81	2.42	1961–71
	(8.38)	(6.14)	(6.14)					
	3.46	−.586	−.293	.137	.069	.83	2.87	1961–71
	(1.68)	(3.41)	(3.41)	(1.15)	(1.15)			

Sources: See Table 2.

[a] The t-statistic values are shown in parentheses; DW is the Durbin-Watson statistic.

hypothesized and the lag distributions were estimated by the Almon Interpolation Technique using a first degree Almon variable to conserve degrees of freedom.

The results of these regressions indicate that vacancy rates and the rate of change of property taxes do significantly affect the rate of change of rents and hence rent levels. In the individual city regressions the vacancy variable was significant at the 95 percent significance level in seven out of ten regressions and had a t-statistic greater than 1.0 in all ten regressions. The poorest results occurred in the Montreal regressions and may partly be explained by disruptions in the housing market that arose from Expo 1967 (the 1967 World's Fair). The impact of variations in property taxes on rents appears less clear-cut in the city regressions since the property tax variable had a t-statistic

Table 2—Regression Results with Pooled Data, Rate of Change of Rents $\dot{R}$, Annual Observations

	Constant	D_1	D_2	D_3	D_4	V	V_{-1}	$\dot{T}$	$\dot{T}_{-1}$	R^2
1)	4.19	.20	1.05	−.02	−.44	−.657				.44
	(7.48)	(.32)	(1.65)	(.04)	(.71)	(5.98)				
2)	4.80	.17	1.61	.01	−.36	−.555	−.278			.61
	(9.51)	(.31)	(2.77)	(.02)	(.63)	(8.02)	(8.02)			
3)	2.43	.74	1.62	.39	.19	−.527		.121	.060	.58
	(3.49)	(1.22)	(2.62)	(.70)	(.31)	(4.88)		(3.76)	(3.76)	
4)	3.40	.55	1.80	.29	.02	−.466	−.233	.087	.044	.68
	(5.09)	(1.04)	(3.34)	(.60)	(.03)	(6.61)	(6.61)	(2.93)	(2.93)	

Sources: R = rental index component of the consumer price index by city, supplied by Statistics Canada. V = vacancy rate in privately initiated apartment structures of 6 or more units, Central Mortgage and Housing Corporation (*CMHC*) 1971, p. 17. Prior to 1963 vacancy rates were obtained from the following rental vacancy surveys: *CMHC*, June 1962; *CMHC*, 1964; and Metropolitan Toronto Planning Board, June 1962. Vacancy rates were adjusted to correspond to the same month. T = index of property taxes, based on data supplied by municipal tax offices.

greater than 1.0 in only four out of five regressions and was significant at the 95 percent significance level in only two regressions. However, both the vacancy and property tax variables were highly significant in the pooled regressions, and these regressions are a more reliable indicator of the underlying relationships than the individual city regressions because of the much larger degrees of freedom.

Two interesting implications also follow directly from our analysis. First, if we define the natural vacancy rate (the rate at which there is no excess demand nor supply) as that rate at which rents are in equilibrium, then from equation 4), Table 2 (the statistically "best" equation), we get a natural vacancy rate of approximately 5.0 percent for Toronto, Winnipeg, and Vancouver, of 5.6 percent for Halifax, and of 7.4 percent for Montreal.[8] Second, if we assume that approximately 22 percent of gross rents go to property taxes (as in Toronto), then equation 4), Table 2, indicates that landlords are able to pass approximately 60 percent of an increase in property taxes on to the tenant since a 10 percent increase in property taxes would increase rents by 1.31 percent.

Therefore, from our analysis we can conclude that excess demand or supply conditions as reflected in the vacancy rate do significantly affect the rate of change of rents, and that the price of housing services is quite responsive to variations in market conditions. Furthermore, it appears likely that landlords are able to pass a significant portion of the increases in their operating costs to the tenant in the form of higher rents.

---

[8] The higher equilibrium vacancy rate in Montreal may partly be explained by Expo 1967 and expectations generated by it, and by the existence of partially segmented English and French speaking markets.

REFERENCES

S. Almon, "The Distributed Lag Between Capital Appropriations and Expenditures," *Econometrica,* Jan. 1965, *33,* 178–96.

D. M. Blank and L. Winnick, "The Structure of the Housing Market," *Quart. J. Econ.,* May 1953, *67,* 181–208.

F. de Leeuw and N. F. Ekanem, "Time Lags in the Rental Housing Market," *Urban Stud.,* Feb. 1973, *10,* 39–68.

——— and ———, "The Supply of Rental Housing," *Amer. Econ. Rev.,* Dec. 1971, *61,* 806–17.

A. Downs, "Moving Toward Realistic Housing Goals," in K. Gordon, ed., *Agenda for the Nation,* Washington 1968.

R. C. Fair, "Disequilibrium in Housing Models," *J. Finance,* May 1972, *27,* 207–21.

S. Maisel, "A Theory of Fluctuations in Residential Construction Starts," *Amer. Econ. Rev.,* June 1963, *53,* 359–83.

L. B. Smith, "A Model of the Canadian Housing and Mortgage Markets," *J. Polit. Econ.,* Oct. 1969, *77,* 795–816.

Central Mortgage and Housing Corporation, *Canadian Housing Statistics,* Ottawa 1971.

———, *Rental Vacancy Survey: Greater Montreal Area,* Montreal 1962.

———, *Greater Vancouver Vancancy Survey,* Vancouver 1964.

Metropolitan Toronto Planning Board, *Vacancy Rates in Apartments: Metropolitan Toronto,* Toronto 1962.

Phase II, Cost-Push Inflation, and Relative Wages

By MICHAEL L. WACHTER*

Much of the recent debate over Phases I and II of the President's Stabilization Program has been conducted by macro economists using aggregate data to support their models and conclusions. The main point of agreement among the participants is that the wage or price equation, that is, the Phillips curve, has shifted adversely so that a given rate of unemployment is concomitant with a higher rate of inflation.[1] Of specific concern is the period from 1969 to 1971 when an unemployment rate which climbed to 6 percent was accompanied by a continuing high rate of wage and price change. Those who support an incomes policy tend to believe that the large oligopolistic, unionized industries have been a primary cause of the shift in the Phillips curve by initiating cost-push inflation. For example, Arthur Okun has argued that an incomes policy "... would demonstrate that the nation will not blithely accept cost-push inflation. The failure of prices and wages to decelerate in 1970–71 cannot be reconciled with any notion that competitive market forces prevail in determining the price level" (p. 498). Support for the cost-push explanation has also been expressed by, among others, Gardner Ackley, Gottfried Haberler, and Lloyd Ulman.[2]

When cost-push pressures are the problem, the main function of an incomes policy is to place restrictions on the rate of wage and price increase in the high wage, noncompetitive sector during periods of slack demand. This discrimination between sectors arises because the noncompetitive sector, virtually by definition, is the home of cost-push pressures. There is little support for a more comprehensive set of controls which would operate during tight labor markets and would restrict the competitive sector as well. In this context, the wage-price freeze of Phases I and II and the Guideposts policy of the early 1960's are essentially similar—the major difference is that the current policy has elements of legal compulsion while the latter was based on moral suasion.

The above argument implies that for an incomes policy to be both needed and successful, one should be able to observe variations in the relative wages between noncompetitive and competitive industries. Specifically, the wage structure should widen during cost-push inflation as the noncompetitive, high wage industries increase their wage premium; then it should narrow upon the imposition of a Phase II type incomes policy. The major case where this would not be true is where there are 100 percent spillovers from the noncompetitive to the competitive sector; every action taken by the former is exactly matched by the latter. With a complete spillover, the incomes policy could be successful in lowering the aggregate rate of wage or price change, but leave relative wages alone. The overwhelming empirical evidence suggests that although spillover effects may be present, they are far from 100 percent.[3] Hence, to observe cost-push inflation and to justify an incomes policy aimed at the noncompetitive industries, one must be able to observe changes in the interindustry wage structure.

---

* Associate professor, University of Pennsylvania. This paper was written at the time when Phase II was near its end but before Phase III was announced. Financial support for this paper was provided by the Fels Center of Government and the National Science Foundation. The views expressed are my own and do not necessarily represent those of the sponsoring agencies.

[1] There may be some terminological confusion over what is meant by a shift in the Phillips curve. By the Phillips curve, this paper refers to the reduced form tradeoff between wage changes and unemployment. The wage equation, however, may contain other variables, such as the expected rate of inflation or a measure of cost-push pressure, as well as the unemployment rate. A change in the expected rate of inflation, for example, would not shift the wage equation, but it would shift the Phillips curve or tradeoff between wage changes and unemployment.

[2] Haberler, although expressing support for the cost-push explanation, suggests that the appropriate policy response is to dismantle the noncompetitive elements in the economy.

Most of the recent research on incomes policies has studied the aggregate rate of wage change. A popular approach is to insert a dummy variable into the wage equation, which takes on the value of unity when the policy is in effect and zero elsewhere. A significant negative sign indicates that the policy has been successful in restraining inflation. There are three major problems with this approach.[4] First, since a dummy variable is not precise, it can represent a number of conflicting hypotheses. Second, there is a problem of causality. Even if an incomes policy is unsuccessful, its dummy variables may have the correct and significant sign. Since, at least in the United States, the government is likely to abandon an incomes policy or alter its framework as soon as it appears to be unsuccessful, the dummy variable would be equal to unity (that is, the policy is in effect) only when the policy appeared to be working. An example is provided by the timely demise of the guideposts policy in 1967. In essence, the residuals of the wage equation could be explaining the dummy variable, rather than the reverse. Third, to test for the success of an incomes policy one must specify the mechanism through which the policy operates. If an incomes policy is specifically designed to have a differential impact across industries, then a direct test of the impact of the policy should involve inspection of relative wages.

Some research has been done which specifically tests for the impact of an incomes policy or for the presence of cost-push inflation, by analyzing relative wages.[5] A weakness of much of this work is that it assumes that without an incomes policy, or exogenous outbreaks of cost-push pressure, the interindustry wage structure would be unchanging. For example, George Perry finds that the wage structure in manufacturing narrows during the guideposts period and, in attributing this decline to the influence of the guideposts, he implicitly assumes that the wage structure would otherwise be unchanged. Similarly, Adrian Throop calculates that relative wages widened during the 1950's and he assumes that this is evidence of cost-push inflation (in the sense that the unions formed in the 1930's, thwarted by wartime controls, were unable to establish their desired wage differential until the 1950's).

In earlier papers (1970a, b) the author developed a model which indicates that, independently of government intervention and exogenous cost-push pressures, the interindustry wage structure varies systematically over the business cycle. Using a variant of that model here, I shall present an empirical test of the efficacy of recent incomes policies and of the existence of cost-push inflation in the 1969–1972 period.

Section I provides a brief summary statement of the model[6] and Section II analyzes the empirical evidence in the context of a relative wage equation. Implications are discussed in a concluding section.

I. The Relative Wage Model

For expositional purposes, the economy may be divided into two sectors, a competitive and a noncompetitive sector. More generally, of course, industries differ along a continuum in the degree of noncompetitive behavior exhibited in their product and/or labor markets. The distinguishing features of the competitive industries are that they adjust rapidly to market conditions and pay a wage relatively close to the market clearing wage.[7] Noncompetitive industries, on the other hand, pay a wage premium above the market clearing wage and change their wage infrequently.[8] Hereafter, the desired wage premium of the noncompetitive sector is denoted by Ψ and the period over which wages are inflexible is referred to as the "contract period." In a sense, the noncompetitive sector may be viewed as first setting its wage as a markup on the competitive wage and then, in order to preserve this differential over the contract period, adding the expected rate of wage inflation. Variation over time in relative wages are thus due to two elements: changes in the desired wage premium of the high wage sector and differences between the actual and the expected rate of wage change. Hence

---

[3] For tests of the spillover model see, for example, the papers by Otto Eckstein and Thomas Wilson, Gail Pierson, Timothy McGuire and Leonard Rapping, and Sherwin Rosen. These studies indicate that a spillover effect is present, but is less than 100 percent.

[4] See, for example, George Perry's defense of the guideposts policy. Studies which utilize simulation techniques have comparable problems.

[5] Although this study concentrates on the relative wage question, an analogous study could be done with relative prices as the dependent variable.

[6] Aspects of the model described below are taken from my 1970a paper and the paper by Stephen Ross and myself. The interested reader is referred to the earlier papers for a more detailed discussion of the model.

$$(1) \qquad CV_t = f_1[Dw_t^e, Dw_t, \Psi_t]$$

where CV is the coefficient of variation among interindustry wages, Dw^e is the expected percentage rate of wage change, and Dw is the actual rate of wage change.

The impact of expectational errors on CV_t is included in the relative wage equation in a traditional manner. It is assumed that the expectation of the wage change over the contract period is formed as a function of those variables which presumably determine the actual rate of wage change over the contract period; namely, the (inverse) expected unemployment rate $(U^e)^{-1}$, and the expected rate of inflation, DP^e. The variables $(U^e)^{-1}$ and DP^e, in turn, are assumed to be formed adaptively on current and past levels of U^{-1} and DP. Hence, relative wages are dependent on a distributed lag of U^{-1} and DP; that is,

$$(2) \quad CV_t = f_2\left[\sum_i \omega_i DP_{t-i}, \sum_i \nu_i U_{t-i}^{-1}, \Psi\right]$$

An increase in DP or U^{-1} above their expected values should cause a narrowing of the wage structure as the rate of wage change in the high wage sector falls behind the rate in the competitive sector; thus, for low values of i, ω_i and ν_i should be negative. If expectations fully adjust over time to the actual rate of wage change, the sum of the lag weights would equal zero, so that for later values of i, ω_i and ν_i should be positive. In this case, the long-run equilibrium wage structure is solely dependent upon Ψ.

The second major determinant of relative wages—changes in the wage premium—is somewhat more difficult to represent quantitatively. In a popular version of the cost-push argument, Ψ may be interpreted as increasing exogenously in response to growing union militancy. Consequently, Ψ behaves as an unspecified shift parameter in equation (2).[9] On the other hand, in the strict neoclassical literature, changes in the desired wage premium are dependent on the strength of the basic noncompetitive elements in the product and labor markets. Since concentration ratios and degree of unionization are largely constant or vary in a trend-like manner over the postwar period, changes in the desired wage premium can simply be represented by a trend variable; that is, $\Psi = \Psi(T)$.

---

[7] The assumption that the competitive sector pays the market clearing wage can be modified to allow for spillover effects. Models which include a spillover between the competitive and noncompetitive sectors appear in Rosen, and Orley Ashenfelter, George Johnson, and John Pencavel. Their work suggests that, under fairly general assumptions, spillovers affect the level, but not the existence of an equilibrium interindustry wage structure.

[8] For firms with labor unions, the above characterization of the wage position for the noncompetitive firm is obvious. The successful union achieves a wage above the market clearing wage and the length of time over which the wage strategy is unchanged is equal to the length of the collective bargaining contract. Elsewhere it has been argued that oligopolies, even in the absence of labor unions, may desire to pay a wage premium over the market clearing wage and to alter that wage only at discrete intervals (see, for example, the paper by Ross and myself). The basis of the argument is that such a wage policy makes it easier for the administered firm to maintain its oligopoly pricing strategy.

[9] Attempts to capture the socio-political forces that determine the wage premium, in the context of a wage equation, have been made by Ashenfelter, Johnson, and Pencavel and A. G. Hines.

In a more general view of the neoclassical model, unions and oligopolies may have utility functions containing a number of goals and, in response to changing economic conditions, they may move systematically to different points on their indifference map placing more or less stress on the wage premium relative to other goals. Furthermore, the bargaining power of the parties may vary cyclically. This can occur even though the underlying (secular) strength of noncompetitive elements remain unchanged. The diverse literature on union goals and bargaining models (see, for example, Allan Cartter, John Dunlop, Charles Holt, and Albert Rees) suggests that both current and lagged values of the unemployment rate and possibly the rate of inflation are important determinants of cyclical or short-run changes in the desired wage premium. Unfortunately, a priori theorizing concerning the form of the wage premium function in the bargaining context, has proven difficult. Even the signs of the first derivatives are largely empirical questions. Since in this paper relative wages are already a function of current and lagged values of the unemployment rate and the rate of inflation, the fact that the wage premium is responsive to these variables adds little to the general form of equation (2). This is especially true if expectational or adjustment effects cause the desired premium to respond with a lag to the observed unemployment and inflation rates. A consequence of the unspecified nature of the wage premium effect, however, is that it is impossible to separately identify the wage premium and expectational effect parameters. Furthermore, there is no longer a presumption that the sum of the weights $\Sigma_i\omega_i$ and $\Sigma_i\nu_i$ should equal zero. Hence, the equilibrium wage structure may be a function of the unemployment rate and the rate of inflation.

II. Empirical Results

The relative wage equation is estimated using two data series on interindustry wages as the dependent variable: first, the coefficient of variation among two-digit manufacturing industries for annual data, 1947–71 and second, the coefficient of variation among the approximately seventy industry headings used by the Department of Commerce in reporting national income data, for annual data 1929–1958. The latter series was constructed by H. Gregg Lewis.[10] The estimated equation is of the form

$$(3) \qquad CV_t = \alpha_0 + \alpha_1 \sum_{i=0}^{n} \omega_i DP_{t-i} + \alpha_2 \sum_{i=0}^{m} \nu_i U_{t-i}^{-1} + \alpha_3 T_t + \alpha_4 X_t$$

where CV_t is the coefficient of variation, T is a time trend representing secular changes in the strength of noncompetitive forces, and X is a vector of dummy variables representing periods of wage-price controls. One would expect the lags in equation (3) to be quite long. First, to the extent that wage contracts in the unionized sector are as long as three years, the adjustment lag of noncompetitive firms can be expected to be at least three years. Second, in general, the longer the time horizon of the forecast (made necessary by the fixed wage contract), the longer the expectational lag. Furthermore, the above discussion has assumed that the competitive sector adjusts immediately. Lags in that sector, at least in part, add to the length of the lags in the relative wage equation.[11]

The results are presented in Tables 1 and 2. In each equation dummy variables are introduced wherever relevant, in terms of

---

[10] The coefficient of variation is a more appropriate measure of relative wages than is a variable of the form w^{nc}/w^c, where nc represents noncompetitive and c represents competitive industries. The latter would require an a priori specification of which industries fall into each category. If a simple average were then used, it would assume that all industries were equally, say, noncompetitive in the nc category. A division of industries into competitive and noncompetitive categories, using the above model, is made in my 1970b paper. To use that classification here, however, would obviously bias the results in favor of the model. Although the underlying wage data are corrected for overtime, they are uncorrected for differences in skill mix across industries. In my 1970a paper, it is shown that high wage industries have relatively skilled labor forces but also pay relatively high wages for given skill categories.

[11] This is discussed more fully in the paper by Ross and the author.

TABLE 1—RELATIVE WAGES 1947–71

	Constant[a]	U^{-1} [b]	DP[b]	Trend	Guide-posts	Korean War	$R^2/D.W.$[c]
i	0.2386	−0.2704	−0.5283				.9487
	(43.3)	(9.9)	(7.4)				1.94
ii	0.2339	−0.2618	−0.3326			−0.0119	.9617
	(49.7)	(11.4)	(5.2)			(6.55)	1.50
iii	0.2391	−0.2807	−0.3288		−0.0039	−0.0206	.9651
	(43.8)	(11.4)	(5.4)		(1.7)	(7.0)	1.88
iv	0.2288	−0.2636	−0.3494	0.0187	−0.0035	−0.0184	.9658
	(22.0)	(9.3)	(5.5)	(1.2)	(1.5)	(5.3)	1.82

Note: The dependent variable is the coefficient of variation of average hourly earnings excluding overtime in the two-digit manufacturing industries.

[a] The terms in parenthesis are the *t*-statistics of the coefficients.

[b] The length of lags for both *DP* and U^{-1} are five years long in all equations and the lags are constrained to have a zero weight in $t-5$. The coefficients presented for U^{-1} and *DP* are the sum of the lag weights.

[c] The $R^2/D.W.$ presents the adjusted R^2 on the top line and the Durbin-Watson static on the second line.

the time period of the equation, to control for the impact of wage regulation during World War II, the Korean War, and the guidepost period. The discussions of the results focuses on Table 1; the Lewis data are utilized largely to indicate that the model is relevant to a broader set of industries and years than the postwar manufacturing series. This paper's concentration on the manufacturing sector is not restrictive, since both Phase II and the guideposts of the 1960's were aimed, in large part, at the noncompetitive industries in this sector. Of the two other sectors which have often been singled out for special mention as sources of cost-push inflation, the construction and the municipal employees sector, the former has been covered separately, prior to Phase I, by the Construction Industry Stabilization Committee, while the latter has only been indirectly influenced by the controls program.

In equation (i) in Table 1, the inflation and unemployment terms are included and

TABLE 2—RELATIVE WAGES 1929–58

	Constant[a]	U^{-1} [b]	DP[b]	Korean War	Trend	World War II	$R^2/D.W.$[c]
i	0.3331	−0.1168	−0.4912				.8910
	(79.4)	(5.2)	(6.7)				0.94
ii	0.3329	−0.1312	−0.4916			0.0215	.9120
	(88.3)	(6.3)	(7.4)			(2.6)	0.83
iii	0.3075	−0.1566	−0.5718		0.0838	0.0334	.9390
	(37.4)	(8.2)	(9.5)		(3.3)	(4.3)	1.47
iv	0.3084	−0.1596	−0.5398	−0.0101	0.0361	0.0289	.9463
	(39.9)	(8.9)	(9.2)	(2.0)	(3.7)	(3.8)	1.60

Note: The dependent variable is the coefficient of variation of average hourly earnings in the seventy industry headings used to calculate the national income accounts. See H. Gregg Lewis.

a,b,c: See Table 1.

the institutional variables are excluded.[12] The (long-run) coefficients on U^{-1} and DP are significant and indicate that a decrease in unemployment and an increase in inflation both act to narrow the wage structure. The equation's $\overline{R}^2$ is .9487 and the Durbin-Watson is 1.94.[13,14] The fit of the equation supports the view that the wage structure is a dynamic factor over the cycle. It is thus inappropriate to interpret any widening of the wage structure as evidence of an exogenous outbreak of cost-push inflation. Nor is it appropriate to suggest that any narrowing of the wage structure that occurs during a period when an incomes policy is operating is evidence of the success of that policy. The lag structures on DP and U are complex in that they represent both the formation of expectations in the noncompetitive sector and adjustment lags in the competitive and the noncompetitive sectors. If only expectational effects were present and the wage premium were not a function of DP and/or U^{-1}, it might be expected that the lag weights should sum to zero. In that case a constant rate of inflation or unemployment is eventually anticipated by the high wage, noncompetitive sector so that the wage structure in steady state returns to an equilibrium value. In Table 1, however, the sum of the weights for DP and U^{-1} are negative. This indicates that for each given rate of inflation and rate of unemployment there exists a different long-run wage structure. This supports the view that the desired wage premium, Ψ, is dependent on DP and U. It should be emphasized, however, that a continuing high inflation rate, or low unemployment rate only has a once-and-for-all narrowing effect on the wage structure.

Equation (iv) of Table 1 includes the institutional variables: a trend to represent secular changes in the strength of noncompetitive elements over the postwar period and dummy variables that take on the value of unity during the Korean War and guideposts period, respectively. The trend is insignificant and is omitted in equation (iii). The Korean War and guideposts variables both indicate an exogenous narrowing of the wage structure due to the impact of government controls. The guideposts variable, however, is significant only at confidence levels below 10 percent (on the two-tailed T-test).[15]

The guideposts were in effect beginning in 1962 and ending in 1966. The coefficient of variation narrowed by .019 during the period immediately surrounding the guideposts. According to equation (iv) of Table 1, only one-sixth of that decline should be attributed to the effects of the guideposts. Even this small impact, however, is of a short-run nature. After 1966 (the guideposts dummy variable is equal to zero) and the wage structure returns to the level that would have existed if the policy had never been in effect. Furthermore, if one believes that the controls *alone* were responsible for narrowing the wage structure during the 1962 through 1967 period, then one might expect the wage structure to widen again after the controls were lifted. In fact, the reverse occurs; as labor markets continue to tighten after the lifting of the guideposts, the wage structure continues to narrow. Indeed, the rate of decline in the wage premium of the high wage sector is greater after the incomes policy is lifted than during the years when the policy is being enforced.

---

[12] Several different Almon lags were tried and the results, including the shape of the lag and the significance of the lag weights, were insensitive to small alterations in the length of the lag, the degree of the polynomial, and whether or not the final lag weight was constrained to be zero. To ease comparisons, all lags for U^{-1} and DP in Tables 1 and 2 have weights along a second degree polynomial and have the weight at $t-5$ constrained to be zero.

[13] No attempt is made here to compare directly the results of Table 1 with those in my 1970a study. The equations in that paper, which were estimated for the 1947–67 period, were specified somewhat differently from equation (3). The two sets of results, however, are compatible and, in fact, the equation in the earlier paper can forecast the direction of change of the coefficient of variation in the 1968-71 period.

[14] The lag weights for DP are -0.0252, -0.0379, -0.1145, -0.1823, and -0.1683. The lag weights for U^{-1} are -0.1300, -0.0562, -0.0302, -0.0280, and -0.0259. Hence the lag to 50 percent completion is in year $t-3$ for DP and is in year $t-1$ for U^{-1}. The pattern of the weights for DP and U^{-1} is largely the same in the other equations in Table 1 so that they are not discussed.

[15] One might claim that the guideposts variable would be more significant if it were not represented by a dummy variable. Clearly, the impact of the guideposts varied over the 1962–66 period. The use of a dummy variable, however, is justified given the use of a similar dummy variable to represent the guideposts' effect in Phillips curve equations.

For the years of the hypothesized cost-push inflation, 1970–71, the wage structure widens by .009. Equation (iv) of Table 1, which assumes that only normal cyclical influences are present, explains the widening with residuals that are smaller than the standard error of the equation. Furthermore, the residual is negative for 1970, indicating that the wage premium of the high wage sector is even lower than expected on the basis of past behavior. Hence, the relatively large increases in the wages for noncompetitive manufacturing industries in the 1970–71 period is a repeat of the historical tendency of the wage structure to widen during slowdowns in economic activity. The relative wage equations of Table 1 explain both the direction and magnitude of this development.

The annual data mask some interesting quarterly variation in relative wages which are worth analyzing. President Nixon's incomes policy began in the third quarter of 1971. One might therefore expect the wage structure to narrow after that date. The data indicate, however, that the wage structure continued to widen throughout 1972. Although this widening of relative wages may be due to wage contracts agreed upon but not implemented before Phase I began, there is no evidence in this data to suggest that Phase I or Phase II has had the anticipated impact of narrowing relative wages. In fact, a preliminary examination of residuals from the equations of Table 1 suggest that relative wages are *narrower* than expected for the quarters immediately preceding the introduction of controls, but are *wider* than expected after the summer of 1971. This finding also contradicts an often made suggestion that wage controls have a beneficial impact on allocational efficiency by reducing the union-nonunion differential, thus approximating the competitive market solution. To the extent that the Phase II controls have had any impact on relative wages, it has been in the direction of widening differentials.

Although the initial widening of the wage structure after controls are introduced is not the anticipated result, it could be explained by an alternative view of the impact of controls. In this explanation Phase II has its greatest impact by lowering inflationary expectations rather than by enforcing lower wage settlements in the high wage sector. Assume that expectational factors enter into all wage bargains, whether competitive or noncompetitive. Then, since the relatively long wage contracts of the noncompetitive sector ensure that inflationary expectations in that sector adjust more slowly than do expectations in the more rapidly adjusting low wage sector, the Phase II incomes policy could have the very short-run impact of widening the wage structure.[16]

The model estimated on the Lewis data uses a broader group of industries than the manufacturing group and an overlapping, but different time period. It provides additional support for the relative wage model.[17] The rate of inflation and the unemployment variable both enter the equations with significant and negative coefficients. A comparison of the coefficients between Tables 1 and 2 indicates a relatively larger effect for unemployment and a considerably smaller role for inflation in determining interindustry differentials in the postwar manufacturing data than in Lewis' 1929–58 series. Lewis himself stresses unanticipated inflation as the cause of cyclical variation in interindustry differentials and this is largely supported by an analysis of the 1929–58 data set. Differences in the dependent variable, however, makes speculative any explanation of differences in the coefficients. The significance of the dummy variables for World War II and the Korean War indicates that the types of regulation adopted during the wars did have a short-run impact on relative wages. The positive coefficient for the World War II dummy suggests that the government prevented the wage structure from narrowing as much as was dictated by market conditions. This may be explained by the fact that when the government moved to regulate wages after the war began, it acted to freeze a wage structure that was relatively wide as a consequence of the Great Depression. Also of interest is the positive and significant trend variable. This may reflect the growth of labor unions in the manufacturing sector over the 1929–58 period.

---

[16] This result is especially true where the incomes policy is expected to be short-lived. As mentioned above, the longer the period of wage inflexibility, the longer the time horizon of the expectation and hence, the smaller the effect on expectations of a temporary incomes policy. If the incomes policy is expected to be permanent the result will still hold, but only because of the relatively long adjustment lag in the noncompetitive sector due to the length of the wage contract.

[17] The Lewis data are from the author's classic work on the impact of labor unions. The model presented in this paper draws upon his findings.

III. Policy Implications

The basic point of this paper is that cost-push inflation has not been the source of the 1970–71 inflation. The observation that wages have been rising in the unionized, oligopolistic industries faster than in the competitive industries is a repeat performance of the historical tendency for the wage structure to widen when the unemployment rate is high. Furthermore, the magnitude of the rise in the wage premium of the high wage sector, prior to the introduction of controls, is not greater than in previous periods of comparable economic conditions. Ironically, it is only after controls are introduced that relative wages widen beyond their traditional level, that is, the residuals of the equations of Table 1 turn positive.

Hence, the results of the relative wage model suggest that if one purpose of controls has been to lower relative wages in the highly unionized manufacturing industries, it has not succeeded. This lack of impact on relative wages by Phases I and II may, however, be a saving feature. That is, in the absence of cost-push inflation in the precontrols period, this type of policy does not seem to have been needed.[18] One could even argue that Phases I and II were, in part, "successful" because they did not artificially alter the wage structure.

This paper does not imply that cost-push pressure did not exist anywhere in the economy in the precontrol period, 1970–71. The evidence presented in this paper, however, suggests that although cost-push inflation may have existed in some areas of the economy, it was not an economy-wide problem. Where individual sectors pose an inflation problem—the construction and municipal employee sectors are often mentioned in this respect—perhaps the appropriate government response is specific policies aimed at the offending sector(s).[19] Such policies could involve the removal of past government restrictions on competition as well as some control over the relevant wages and prices.[20]

Recently, it has been suggested that controls be maintained even in tight labor markets. The justification for such a policy would not be based on the presence of cost-push inflation, but rather on the ability of controls to lower inflationary explanations. Such a policy, however, would have two major problems. First, although individuals might today reduce their expectations when a controls policy is introduced, they would soon learn that such a policy cannot lower long-run inflation rates without appropriate monetary and fiscal policy. In both the guideposts experience and more recently with Phases I and II, controls have been accompanied by highly expansionary aggregate demand policies. Hence, even if wage (and price) increases are effectively restrained by law, expectations will eventually adjust to the realities caused by monetary and fiscal policy and the major consequence of controls will be suppressed inflation. Second, the existence of a highly dynamic interindustry wage structure will pose problems for a policy aimed at overall inflationary expectations as well as for one oriented towards the noncompetitive sector. Popular and easily enforced across-the-board rules will cause serious allocational effects. A controls policy that would have a smaller disruptive effect on the economy might include two provisions; it could recognize the normal cyclical variation in relative wages, thus avoiding across-the-board wage rules, and it could operate directly through the market mechanism by altering the desired wage premium sought by the unions and the oligopolies.[21] The problems with such a policy, however, are that the controls would be both difficult to establish and to enforce. Clearly, when relative wages (and prices) vary significantly in the short run there is no simple replacement for the wage (and price) mechanism.

---

[18] The cost of an inappropriate program of wage and price controls is discussed by Milton Friedman. One problem involves the miseducation of the public. The notion that inflation can be 'cured" by controls does not encourage the appropriate use either of monetary and fiscal policy to deal with overall problems of excess demand or of structural remedies to relieve specific bottlenecks in affected sectors.

[19] Average hourly earnings have been increasing in the construction industry relative to both the low wage industries and the other high wage industries over the entire postwar period. The construction industry however is sufficiently unique and complex that no generalizations are possible until a study disaggregated by occupation and region is available.

[20] Although this paper has discussed aspects of cost-push inflation and incomes policies it has done so from the perspective of relative wages. Hence although the paper has discounted the likelihood that the current inflation is due to exogenous cost-push elements, it has not attempted to explain what is causing the inflation. There are a number of potential explanations of the current inflation; the one which is most compatible with the relative wage model is discussed in a recent paper by Ross and the author.

---

[21] A detailed discussion of such a policy is beyond the scope of the present paper. The notion that wage rules should not be applied without reference to relative wage structures was an important component of John Dunlop's approach on the construction wage stabilization board. An incomes policy which operates through a market mechanism has been explored in detail by Henry C. Wallich and Sidney Weintraub. See also Dunlop (1966).

REFERENCES

G. Ackley, "An Incomes Policy for the 1970's," *Rev. Econ. Statist.*, Aug. 1972, *54*, 218–23.

O. C. Ashenfelter, G. E. Johnson, and J. H. Pencavel, "Trade Unions and the Rate of Change of Money Wages in United States Manufacturing Industry," *Rev. Econ. Stud.*, Jan. 1972, *39*, 27–54.

A. M. Cartter, *Theory of Wages and Employment*, Homewood 1959.

J. T. Dunlop, "Guideposts, Wages and Collective Bargaining," in G. P. Shultz and R. Aliber, eds., *Guidelines, Informal Controls and the Market Place*, Chicago 1966, 81–96.

———, *Wage Determination Under Trade Unions*. New York 1944.

O. Eckstein and T. A. Wilson, "Determination of Money Wages in American Industry," *Quart. J. Econ.*, Aug. 1962, *70*, 379–414.

M. Friedman, "The Case Against the Guideposts," in G. P. Shultz and R. Z. Aliber, eds., *Guidelines, Informal Controls and the Market Place*, Chicago 1966, 17–61.

G. Haberler, "Incomes Policy and Inflation: Some Further Reflections," *Amer. Econ. Rev. Proc.*, May 1972, *62*, 234–41.

A. G. Hines, "Trade Unions and Wage Inflation in the United Kingdom, 1893–1961," *Rev. Econ. Stud.*, Oct. 1964, *31*, 221–52.

C. C. Holt, "Job Search, Phillips' Wage Influence and Union Influence: Theory and Evidence," in E. S. Phelps et al., eds., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970, 53–123.

H. G. Lewis, *Unionism and Relative Wages in the United States*, Chicago 1963.

T. W. McGuire and L. A. Rapping, "The Role of Market Variables and Key Bargains in the Manufacturing Wage Determination Process," *J. Polit. Econ.*, Sept. 1968, *76*, 1015–36

A. M. Okun, "The Mirage of Steady Inflation," *Brookings Papers*, Washington 1971, *2*, 485–93.

G. L. Perry, "Wages and the Guideposts," *Amer. Econ. Rev.*, Sept. 1967, *57*, 897–904.

G. Pierson "Union Strength and the U.S. 'Phillips Curve'," *Amer. Econ. Rev.*, June 1968, *58*, 456–67.

A. Rees, *The Economics of Trade Unions*, Chicago 1962.

S. Rosen, "Trade Union Power, Threat Effects and the Extent of Organization," *Rev. Econ. Stud.*, Apr. 1969, *36*, 185–96.

S. A. Ross and M. L. Wachter, "Wage Determination, Inflation, and the Industrial Structure," *Amer. Econ. Rev.*, Sept. 1973, *63*, 675–92.

A. W. Throop, "The Union-Nonunion Wage Differential and Cost-Push Inflation," *Amer.*

Econ. Rev., Mar. 1968, *58*, 79–99.

L. Ulman, "Cost-Push and Some Policy Alternatives, *Amer. Econ. Rev. Proc.*, May 1972, *62*, 242–50.

M. L. Wachter, (1970a) "Cyclical Variation in the Interindustry Wage Structure," *Amer. Econ. Rev.*, Mar. 1970, *60*, 75–84.

———, (1970b) "Relative Wage Equations for U.S. Manufacturing Industries, 1947–1967," *Rev. Econ. Statist.*, Nov. 1970, *52*, 405–10.

H. C. Wallich and S. Weintraub, "A Tax-Based Incomes Policy," *J. Econ. Issues*, June 1971, *5*, 1–19.

Earnings, Employment, and Racial Discrimination: Additional Evidence

By B. F. KIKER AND W. PIERCE LILES*

In a recent article (1970) in this *Review*, James Gwartney uses census data and data from the *Coleman Report* to decompose the gross income differentials between whites and blacks into two categories: a difference attributed directly to productivity (including locational) factors, and a residual unaccounted for by productivity factors, which could be attributed to racial discrimination in income. According to two of Gwartney's critics, Orley Ashenfelter and Michael Taussig, perhaps the most interesting result of Gwartney's computations is that differences in scholastic achievement levels explain a large portion of the difference in black/white incomes. This result is derived by adjusting years of schooling taken from census data using the scholastic achievement scores given in the *Coleman Report* (see Gwartney, 1970, p. 400).

Ashenfelter and Taussig argue, among other things, that achievement differentials for students at various grade levels (6, 9, and 12 in 1965) taken from the *Coleman Report* do not accurately reflect the actual achievement differentials for those who completed comparable grade levels in the labor force in 1959. They assert (and Gwartney concurs) that "Ideally this [Gwartney's] approach calls for a longitudinal study of a sample of representative individuals in order to obtain joint observations on lifetime earnings streams, test scores, years of schooling . . . " (p. 747), other productivity-enhancing variables, and perhaps cultural and environmental factors that affect test scores. Although longitudinal data were unavailable to both Gwartney and Ashenfelter and Taussig, we have such data that include many of the variables that Ashenfelter and Taussig suggest.

The objective of this paper is to furnish additional evidence on the effect of racial discrimination on earnings and employment. Although our methodology is similar to Gwartney's,[1] our data allow for the *simultaneous consideration* of the effect of years of schooling and scholastic achievement on gross black/white earnings and employment differentials. We consider also the influence on the black/white earnings and employment differentials of age, marital status, family size, geographic region, and city size. In addition, the effect of occupation on earnings differentials is analyzed. We are concerned specifically with attempting to answer the questions: 1) What portions of black/white earnings and employment differentials can be attributed to distributional differences in productivity and locational factors between blacks and whites, and what portion may be attributed to racial discrimination? 2) What portion of the earnings differential can be attributed to distributional differences in occupation between blacks and whites? In an attempt to answer these questions, we compute Laspeyres and Paasche indexes of earnings *and* employment differences. A difficulty with Gwartney's analysis is that he considers the effect of racial discrimination on earnings only, while in reality racial discrimination is also effected through the rate of unemployment.[2]

The major conclusion of the study, that scholastic achievement is far more important quantitatively than years of schooling in explaining observed black/white earnings and employment differences for the men in our sample, is supported by comparison of the adjusted Laspeyres and Paasche indexes with similar indexes calculated from the results of regression analysis. In addition, we show that the regression procedure used by Gwartney may result in biased estimates.

---

* Professor and assistant professor, respectively, University of South Carolina. We are indebted to James Gwartney, J. M. Ryan, M. J. Foxworth, and referees of this *Review* for helpful comments on an earlier draft of this paper. However, the usual disclaimers hold.

[1] See Ashenfelter and Taussig's criticism of Gwartney's methodology, pp. 749–50, and Gwartney's reply (1971, p. 754).

[2] That is, there is a wage (earnings) effect and an employment effect of racial discrimination. Gwartney's analysis in which census data are used combines the two. Our data allow us to distinguish between the two effects.

I. Data

The primary data used in this study are aggregate data obtained from published and unpublished manpower research reports (as well as some of the computer printouts used as a basis for these reports) compiled by the Office of the Assistant Secretary of Defense for Manpower and Reserve Affairs (1969, 1970a, b).[3] The data were provided through the Department of Defense Post-Service Information File and are based on information extracted from personnel data files maintained on first-term separatees.[4] End of service data which consist of information on race,[5] educational level (years of schooling), age, marital status, family size, military occupation, rank achieved, and Armed Forces Qualification Test (*AFQT*) score were extracted from military records. Post-service data which consist of employment status, earnings, civilian occupation, residence, and schooling-training program participation were obtained through a mail survey initiated ten months after separation from the Army. These data were collected on 304,940 of the 388,000 men who separated during the period July 1, 1968, to June 30, 1969. Twelve percent of the respondents were black and 88 percent were white.

The dependent variables used in the following analysis are annual gross earnings and employment rates.[6] Annual earnings were estimated by the Department of Defense by multiplying gross weekly earnings taken at the time of the survey by 52. The employment variable is simply the percentage of the respondents who were employed at the time of the survey. Only those persons who were unemployed *and seeking work* during the survey period were classified as unemployed. Gross earnings and the employment rate of the black and white samples were quite different during the survey period. Earnings of blacks as a percentage of whites is 88.8. The black/white employment index is 94.0 percent. The independent variables are age, marital status, family size, years of schooling, scholastic achievement (*AFQT*), geographic region, city size, and occupational classification.[7]

---

[3] As far as we know, the computer printouts which were necessary in completing this study have not yet been made available to other analysts. Although we would prefer of course to have individual observations, they are unavailable to us.

[4] Separatees with six or more years of active duty, men who receive administrative discharges for medical reasons or unsuitability, and reservists who enter active service for training purposes are not included in the Post-Service Information File.

[5] The white category includes all races other than black. This dichotomous classification will cause a positive bias in the results of our analysis if discrimination is practiced against other minority groups. This bias should be very small since other races constituted only about 1.0 percent of the sample.

[6] While Gwartney uses income data in his study, we are able to use earnings data, which, of course, is preferred. About 9.0 percent of the separatees were in school or training and unemployed or employed part-time. These persons are excluded from our analysis.

[7] The mean age for whites and blacks is 22.3 and 22.6 years, respectively, with all separatees in the age interval 19 to 28 years. Approximately 70 percent of the blacks in the sample are single, while 65 percent of the whites are unmarried. Approximately 92 percent of the white and 87 percent of the black separatees fall in the 0–1 dependent category. Owing to paucity of data, our "years of schooling" consist of the following categories: non-high school graduate (7–11 years of school); high school graduate (12 years); some college (13–15 years); and college graduate (16–18 years). The mean numbers of years of schooling for blacks and whites are 11.8 and 11.9 years, respectively. A larger percentage of blacks (65) than whites (62) are high school graduates. Four percent of the whites in the sample graduated from college, while only 2 percent of the blacks acquired a college degree. The mean *AFQT* score for blacks and whites is 28 and 55, respectively. Approximately 3 percent of the blacks and less than 1/2 percent of the whites in the sample are in the 0–9 score range on the *AFQT*, while 61 percent of the blacks and 19 percent of the whites are in the 10–30 range. An examinee's *AFQT* score, according to B. D. Karpinos, depends on several factors: ". . . on his innate ability; on the level of his educational attainment; on the quality of his education (quality of the school facilities); and on the knowledge he gained from his educational training, or

The restrictive nature of the sample should be emphasized. Since it consists of males who have been out of the Army ten months, estimates of earnings' potential of either schooling or measured scholastic achievement based on such a short gestation period are perhaps heroic.[8]

II. Analysis

A meaningful analysis of labor-market differences between blacks and whites must be based on an assumed homogeneity between the races with respect to the independent variables. If, in fact, heterogeneity exists, observed differences in earnings and employment could depend less on race than on productivity, locational, and occupational differences between the groups.

Earnings and employment are generally thought to be positively related to age, years of schooling, family size, and scholastic achievement. Married men are usually believed to experience less unemployment and earn more than single men. If differences exist in the distributions of these independent variables between blacks and whites, we would expect the differences to account for a portion of any observed black/white earnings and employment differential. Hence, we should adjust for these distribution differences. In addition, an overrepresentation of blacks in any of the locales or occupations where differences in average mean earnings and/or employment rates exist may account for some of the observed black/white earnings and/or employment differences. If, for example, blacks are overrepresented in the sample from the South, we would expect that adjustment for this distributional difference would decrease the observed black/white earnings differential.

A. *Laspeyres and Paaschè Indexes*

The observed earnings (or employment) index is the product of the adjusted earnings (or employment) index and an earnings (or employment) index of distributional differences. The adjusted earnings (or employment) index is the hypothetical ratio of black to white earnings (or employment) assuming that the two populations in the sample are distributed identically among the productivity, locational, and/or occupational categories. The index of distributional differences is the hypothetical ratio of black to white earnings (or employment) assuming that within each productivity, locational, and/or occupational category the two populations have identical mean earnings (or employment rates).[9] This black/white index will be less than 100 percent if blacks are overrepresented relative to whites in low earnings (or employment) productivity, locational, and/or occupational categories. The mathematical forms of the Laspeyres earnings indexes are

$$\begin{array}{cccc} & \text{(Unadjusted Index)} & \text{(Earnings Index)} & \text{(Distribution Index)} \\ (1) & \dfrac{Y_B D_B}{Y_W D_W}\cdot 100 = & \dfrac{Y_B D_W}{Y_W D_W} \cdot & \dfrac{Y_B D_B}{Y_B D_W}\cdot 100 \end{array}$$

where Y is the mean earnings of those persons within a productivity, locational, and/or occupational category according to race; D denotes the percentage of the sample within the categories according to race; and B and W are subscripts denoting black and white groupings. The substitution of mean employment rates in the formula for earnings gives the Laspeyres employment indexes. Paasche indexes can be deduced similarly. The adjusted Paasche earnings (or employment) index and the index of distributional differences, however, are based on the assumption that both sample populations have

---

otherwise in or out of school" (p. 6). These factors, of course, are neither independent nor free from environmental and cultural influences. Our regional delineation is identical to that employed by the Bureau of Census. There are four major regions and nine subregions. The city size breakdown is identical to Gwartney's (1970, p. 402).

[8] In addition, if the *AFQT* score is taken as a proxy for ability, as has often been the case, since it is measured postschool, the interpretation of the effect of schooling on earnings holding this measure constant is considerably different than if, for example, innate ability were held constant.

[9] For a more detailed discussion of the Laspeyres (Paasche) index of distribution difference, see Gwartney (1970, p. 398).

Table 1—Black/White Earnings and Employment Indexes, Adjusted for Various Determinants of Earnings and Employment Differences
(Percentages)

	Laspeyres		Paasche	
Explanatory Factor	Earnings	Employment	Earnings	Employment
Unadjusted	88.8	94.0	88.8	94.0
Age	88.8	93.0	88.2	95.1
Marital Status	88.9	93.6	89.2	94.4
Family Size	89.3	93.6	89.1	95.4
Years of Schooling	89.7	94.5	89.7	94.1
AFQT	95.7	96.8	94.8	96.0
AFQT–Years of Schooling	96.5	97.2	95.3	96.1
Region	91.3	93.4	91.5	93.9
City Size	88.2	95.0	88.5	94.7
Region–City Size	91.0	94.3	90.5	94.3
Region–Years of Schooling	92.4	93.8	92.4	93.9
City Size–Years of Schooling	89.0	95.5	89.3	94.8
Region–City Size–Years of Schooling	91.9	94.8	91.3	94.4
Occupation	90.6	—	90.3	—
Occupation–Years of Schooling	90.8	—	90.6	—

Source: Estimates are derived from data obtained from the U.S. Department of Defense (1969, 1970a, b).

the same productivity, locational, and/or occupational distribution as blacks. Table 1 gives the Laspeyres and Paasche indexes adjusted for various determinants of earnings and employment differences. Indexes of black/white distributional differences with respect to the various determinants are presented in Table 2.

Age, marital status, and family size are relatively unimportant in accounting for

Table 2—Indexes of Black/White Distributional Differences with Respect to Various Determinants of Earnings and Employment
(Percentages)

	Laspeyres		Paasche	
Explanatory Factor	Earnings	Employment	Earnings	Employment
Age	100.0	101.1	100.6	98.8
Marital Status	99.9	100.4	99.6	99.6
Family Size	99.4	100.4	99.7	98.5
Years of Schooling	99.0	99.5	99.0	99.9
AFQT	92.8	97.1	93.7	97.9
AFQT–Years of Schooling	92.0	96.7	93.2	97.8
Region	97.3	100.6	97.0	100.1
City Size	100.7	98.9	100.3	99.3
Region–City Size	97.6	99.7	98.1	99.7
Region–Years of Schooling	96.1	100.2	96.1	100.1
City Size–Years of Schooling	99.8	98.4	99.4	99.1
Region–City Size–Years of Schooling	96.6	99.1	97.3	99.6
Occupation	98.0	—	98.3	—
Occupation–Years of Schooling	97.8	—	98.0	—

Source: See Table 1.

the observed black/white earnings differentials. The observed Laspeyres (Paasche) earnings index becomes 88.8 (88.2), 88.9 (89.2), and 89.3 (89.1) percent, respectively, when adjusting for distributional differences in age, marital status, and family size. The adjustment for age results in either no change (Laspeyres), or a 5.3 percent increase (Paasche), in the observed black/white differential of 11.2 percent. Adjustment for age, marital status, and family size alters the black/white employment differential by −16.6 (18.3), −6.7 (6.7), and −6.7 (23.3) percent, respectively.

The adjustment for differences in the years of schooling distributions of blacks and whites accounts for 8.0 (8.0) percent of the observed earnings differential and 8.3 (1.7) percent of the employment differential.[10] The relatively small influence of schooling on the observed indexes is contrary to the a priori conjectures of some social scientists.

The most important factor quantitatively in explaining the separatees' observed black/white earnings and employment differences is *AFQT*. After the data are adjusted for *AFQT* distributional differences, the Laspeyres (Paasche) black/white earnings percentage is 95.7 (94.8), while the Laspeyres (Paasche) employment index is 96.8 (96.0) percent. Hence, differences in scholastic achievement (as reflected by the *AFQT* score) explain, in this analytical frame, approximately 62 (54) percent of the unadjusted black/white earnings differential and 47 (33) percent of the employment differential.

The *AFQT* decile and years of schooling distributions *taken together* are significantly different for blacks and whites in the sample. After adjusting the data for these differences the observed Laspeyres (Paasche) earnings and employment indexes increase to 96.5 (95.3) and 97.2 (96.1) percent, respectively. The adjustment for *AFQT* and years of schooling taken together accounts for 69 (58) percent of the observed earnings difference and 53 (35) percent of the employment differential.

Adjusting the data for black/white regional distribution differences results in a Laspeyres (Paasche) earnings index of 91.3 (91.5) percent. The corresponding employment index is 93.4 (93.9) percent. Adjustment for differences in the regional distributions accounts for 22 (24) percent of the unadjusted earnings differential, while the observed employment difference is increased by 10 (1.7) percent. Hence, blacks appear to be underrepresented in regions associated with high unemployment rates.

Adjustment for city size distributional differences results in increasing the unadjusted earnings differential by from 5.4 (Laspeyres) to 2.7 (Paasche) percent, and reducing the employment difference by 17 to 12 percent. The negative percentage change in the earnings indexes tends to refute John Kain and J. D. Mooney's assertion that recent suburbanization of employment opportunities has influenced adversely the earnings of blacks who reside predominantly in central areas of the cities. The positive percentage change in the employment index suggests, however, that the Kain-Mooney effect may operate on the rate of unemployment.

Further adjustment for the simultaneous effect of region and city size; region and years of schooling; city size and years of schooling; and region, city size, and years of schooling decreases the unadjusted black/white earnings differential by 20 (15), 32 (32), 1.8 (4.5), and 28 (22) percent, respectively. These determinants account for 5.0 (5.0), −3.3 (−1.7), 25 (13), and 13 (6.7) percent, respectively, of the employment difference. The unavailability of joint distributions among all the variables prevents additional simultaneous adjustments.

---

[10] It may be informative to compare the earnings and employment indexes in Table 1 adjusted for distributional differences in years of schooling with similar indexes computed using aggregate male population weights instead of sample weights. The latter observed earnings index is 86.3 percent; the adjusted Laspeyres (Paasche) index is 89.9 (89.9) percent. The observed employment index is 91.3, while the adjusted Laspeyres (Paasche) index is 93.1 (92.0) percent. If we consider only census male population weights within the same age grouping as the sample, the earnings indexes increase to 88.4 (observed), 91.3 (Laspeyres), and 89.9 (Paasche) percent. Comparable employment indexes are 92.2, 94.0, and 92.8 percent.

It is generally contended that the way to reduce differentials in earnings is to eliminate disparities in earning capacities. The preceding analysis suggests that if earning capacities of blacks, as reflected by years of schooling and scholastic achievement, were upgraded to the level of whites, blacks would still earn less and be unemployed more on the average than white cohorts. One of the major assumptions underlying this contention, however, is that there is no racial discrimination in occupational choice. Although race can no longer be explicitly stated as a criterion in hiring qualifications, discrimination in occupational choice is still being practiced. Accordingly, we examine the postcivilian occupational distributions of blacks and whites in the sample to ascertain what effect distributional differences might have on the observed black/white earnings differential. In addition, an adjustment is made in the earnings indexes for joint occupation-years of schooling distributional differences. Again, unavailability of joint distributions prevents further simultaneous adjustments.

An assumption underlying analysis of the occupational distributions is that, in a competitive labor market in the absence of racial discrimination, the expected value of relative earnings of blacks to whites in all occupations would be unity (see Gwartney and Kenneth McCaffree). Adjustment for distributional differences in occupation yields a Laspeyres (Paasche) earnings index of 90.6 (90.3) percent, which represents a 16 (13) percent reduction in the observed earnings differential. Extending the analysis to adjust for distributional differences in occupation and years of schooling taken together increases the Laspeyres (Paasche) index to 90.8 (90.6) percent.

Unfortunately, because of the unavailability of joint distributions, our data are inadequate for considering simultaneously the influence on the black/white earnings and employment differentials of all the independent variables. If, however, following Gwartney, we assume that the effects of these variables on the indexes are independent, then the total effect of the independent variables on the unadjusted black/white earnings and employment differences may be estimated by adding to the observed indexes the marginal effect of each single variable and/or combinations of variables. The marginal effect of various determinants of earnings and employment differences is sum-

TABLE 3—MARGINAL EFFECT OF VARIOUS DETERMINANTS OF EARNINGS AND EMPLOYMENT DIFFERENCES
(Percentages)

	Laspeyres		Paasche	
Explanatory Factor	Earnings	Employment	Earnings	Employment
Age	0.0	−1.0	−0.6	1.1
Marital Status	0.1	−0.4	0.4	0.4
Family Size	0.5	−0.4	0.3	1.4
Years of Schooling	0.9	0.5	0.9	0.1
AFQT	6.9	2.8	6.0	2.0
AFQT–Years of Schooling	7.7	3.0	6.5	2.1
Region	2.5	−0.6	2.7	−0.1
City Size	−0.6	1.0	−0.3	0.7
Region–City Size	2.2	0.3	1.7	0.3
Region–Years of Schooling	3.6	−0.2	3.6	−0.1
City Size–Years of Schooling	0.2	1.5	0.5	0.8
Region–City Size—Years of Schooling	3.1	0.8	2.5	0.4
Occupation	1.8	—	1.5	—
Occupation–Years of Schooling	2.0	—	1.8	—

Source: See Table 1.

marized in Table 3. If, for example, we add to the observed earnings index the sum of the marginal effects of our adjustment for the simultaneous effect of *AFQT* and years of schooling and the effect of region, the observed index increases to from 98 (Paasche) to 99 (Laspeyres) percent, thereby explaining, in our theoretical frame, from 79 to 86 percent of the observed earnings differential.

B. *Regression Model*

Since the most important conclusion of the paper is that scholastic achievement (as reflected by *AFQT* score) is far more important than years of schooling in explaining observed black/white earnings and employment differentials, we reinforce our estimates of the effect of differences in scholastic achievement and years of schooling on the earnings and employment experience of blacks and whites in the sample by using a weighted regression model of the form[11]

$$(2) \qquad Y = a_0 + a_1 S + a_2 A$$

where Y denotes mean earnings or employment for blacks or whites, A is a dummy variable for *AFQT* decile, and S represents a dummy variable for years of schooling. The data points are cross-tabulations for blacks and whites of earnings or employment by *AFQT* decile and years of schooling.[12] The results of the regressions are presented in Table 4. Statistical evidence reveals that the coefficients of years of schooling and *AFQT* are significantly different for blacks and whites (see Gregory Chow). Hence, the "pay-off" to schooling and scholastic achievement is a function of race. If the coefficients were equal, the regression planes would be parallel or coincident, with racial discrimination being reflected by the difference in the intercepts.

The coefficients in Table 4 can be used to compute Laspeyres and Paasche indexes that show the portions of the earnings and employment differences that can be attributed to differences in years of schooling; *AFQT* decile; and years of schooling-*AFQT* decile (taken together). These indexes are shown in Table 5.[13] A comparison of these indexes with relevant ones in Table 1 reveals that the results of the regression analysis support our earlier findings.

Gwartney also attempts to reinforce his estimates of the effect of differences in scholastic achievement and years of schooling on the income experience of blacks and whites by using an unweighted regression model of the form:

$$(3) \qquad \log Y = a_0 + a_1 S + a_2 A + a_3 R$$

where Y denotes annual income, S is years of schooling, A represents scholastic achievement, and R is a dummy variable denoting race. The regression of the logarithm of income on race yields a coefficient which is the logarithm of the white/black income ratio after controlling for years of schooling and scholastic achievement. The reciprocal of the antilogarithm of this coefficient is expressed as an adjusted black/white income index. Since Gwartney used grouped data, it was necessary to weight his observations. Failure to do so should bias his results. In addition, and more importantly, Gwartney assumes that the coefficients of the years of schooling and scholastic achievement variables are identical for blacks and whites. The evidence from our regressions, however, does not support this assumption.

Some evidence on the extent of the bias that might result from the assumption of equal black/white coefficients of years of schooling and scholastic achievement can be offered by applying our data to Gwartney's regression model. Equation (3) yields

---

[11] The use of discrete (dummy) variables in the regression model requires that one class be deleted from each category of discrete independent variables. We omitted the following classes in the analysis: the lowest *AFQT* decile, and non-high school graduates.

[12] The observations in each cell are weighted by the square root of the number of persons in each cell. Of course, the use of grouped data in our regressions results in a downward bias in the standard error of estimate and thus biases downward our results.

[13] The indexes are computed in the same manner as are those in Table 1. Gross average earnings for blacks and whites by *AFQT* decile-years of schooling cells are estimated by equation (2).

TABLE 4—LINEAR REGRESSION OF EARNINGS AND EMPLOYMENT ON RACE, *AFQT* DECILE, AND YEARS OF SCHOOLING
(t-ratios in parentheses)

	Earnings		Employment	
Explanatory Factor	White	Black	White	Black
AFQT Decile:				
10–19	147	176	—	—
	(0.90)	(1.50)	—	—
20–29	341	442	2.25	0.87
	(2.00)	(3.62)	(4.68)	(1.11)
30–39	543	456	2.88	2.24
	(3.20)	(3.72)	(6.48)	(2.79)
40–49	641	694	3.00	1.56
	(3.76)	(5.22)	(6.21)	(1.38)
50–59	725	828	3.27	5.57
	(4.26)	(5.91)	(7.05)	(4.20)
60–69	836	803	3.54	3.10
	(4.92)	(5.23)	(7.46)	(1.88)
70–79	846	1,209	3.44	−0.72
	(4.97)	(6.76)	(7.07)	(−0.33)
80–89	1,006	1,093	3.67	5.21
	(5.92)	(5.77)	(7.97)	(2.18)
90–99	1,162	1,193	3.73	0.59
	(6.81)	(3.75)	(7.30)	(0.13)
Educational Level:				
High School Graduate	154	243	3.52	8.90
	(7.09)	(5.51)	(12.28)	(12.88)
Some College	375	552	2.48	9.68
	(12.13)	(7.61)	(6.09)	(8.51)
College Graduate	1,813	1,808	3.25	4.81
	(41.00)	(14.42)	(5.58)	(2.45)
Constant	6,014	5,512	89.60	81.02
	(35.55)	(45.72)	(248.33)	(123.40)
R^2	0.99	0.99	0.99	0.99
Mean	6,950	6,164	94.9	89.2

Source: See Table 1.

TABLE 5—BLACK/WHITE EARNINGS AND EMPLOYMENT INDEXES ADJUSTED FOR VARIOUS DETERMINANTS OF EARNINGS AND EMPLOYMENT DIFFERENCES[a]
(Percentages)

	Laspeyres		Paasche	
Explanatory Factor	Earnings	Employment	Earnings	Employment
Unadjusted	88.7	93.6	88.7	93.6
Years of Schooling	89.7	94.1	89.6	93.7
AFQT	95.5	94.7	94.8	94.7
AFQT–Years of Schooling	96.2	95.1	95.6	94.8

Source: See Table 1.

[a] Calculated from the linear regressions in Table 4.

adjusted black/white earnings and employment indexes of 93.0 and 94.2 percent, respectively. The regressions properly weighted yield earnings and employment indexes of 94.0 and 94.7 percent, respectively. The latter indexes can be usefully compared to the indexes adjusted for difference in *AFQT*-years of schooling in Table 5.

The basic criticism of Gwartney's procedure by Ashenfelter and Taussig is that since Gwartney does not have joint observations on years of schooling, scholastic achievement, and income, his procedure for estimating the coefficient for scholastic achievement is likely to be biased. Gwartney's procedure can be summarized by estimating (implicitly) regression equation (3) without the variable S, but assuming the relationship

$$S = b_0 + b_1 A \tag{4}$$

If (4) is substituted into (3) we obtain:

$$\log Y = (a_0 + a_1 b_0) + (a_1 b_1 + a_2)A + a_3 R \tag{3a}$$

The coefficient for A in this regression would be an unbiased estimator of a_2 only if schooling has no effect on earnings independently of scholastic achievement, i.e., if $a_1 = 0$; or if years of schooling and scholastic achievement are uncorrelated, i.e., if $b_1 = 0$. The evidence from our earnings regressions in Table 4 shows clearly that the first assumption is unrealistic. The second assumption is also unrealistic. Correlation matrixes show that there is correlation between years of schooling and *AFQT* decile. In addition, we calculate from our data Laspeyres and Paasche indexes of earnings from coefficients estimated from the (weighted) regression equation

$$Y = a_0 + a_2 A \tag{5}$$

which implicitly assumes equation (4). If these indexes (98.1 Laspeyres and 97.0 Paasche), which show black/white earnings adjusted for *AFQT* decile differences, are identical to the comparable indexes in Table 5, then $b_1 = 0$. They, of course, are not identical.

III. Summary

The primary objective of this paper has been to show the singular influence of years of schooling and scholastic achievement (as reflected by *AFQT* decile scores) as well as the simultaneous effect of *AFQT* decile and years of schooling on observed black/white earnings and employment differences for young men of military age who qualify for induction into the armed services. Adjustment for differences in years of schooling, *AFQT* decile, and *AFQT* decile and years of schooling taken together results in reducing the difference in the earnings index by 8.0 (8.0), 62 (54), and 69 (58) percent, respectively. This supports Gwartney's contention that the effect of scholastic achievement is quantitatively the single most important determinant of black/white earnings differences—at least for the young men in our sample. The results of regression analysis reinforce the Laspeyres and Paasche indexes. If the models are substantially valid, the policy implications are apparent. Since our data permit the simultaneous adjustment for *AFQT* decile and years of schooling, the Ashenfelter-Taussig criticism of Gwartney's data is answered.

In addition, we investigated the effect on the observed black/white differentials of independent variables such as age, marital status, family size, geographic region, city size, and occupation. Contrary to the a priori conjectures of some social scientists, age and years of schooling do not appear to have a substantial influence on the unadjusted black/white differentials. However, the simultaneous effect of years of schooling and some other determinants is significant in explaining the black/white differentials. The indexes adjusted for distributional differences in region and city size show that blacks in the sample are underrepresented in regions with high unemployment rates; that the overrepresentation of blacks who reside in large cities leads to enhanced earnings, but to lower rates of employment.

REFERENCES

O. Ashenfelter and M. K. Taussig, "Discrimination and Income Differentials: Comment," *Amer. Econ. Rev.*, Sept. 1971, *61,* 746–50.

G. C. Chow, "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica,* July 1960, *28,* 591–605.

J. Gwartney, "Discrimination and Income Differentials," *Amer. Econ. Rev.,* June 1970, *60,* 396–408.

———, "Discrimination and Income Differentials: Reply," *Amer. Econ. Rev.,* Sept. 1971, *61,* 751–55.

——— and K. M. McCaffree, "Variance in Discrimination Among Occupations," *Southern Econ. J.,* Oct. 1971, *38,* 141–55.

J. F. Kain, "Housing Segregation, Negro Employment, and Metropolitan Decentralization," *Quart. J. Econ.,* May 1968, *82,* 175–97.

B. D. Karpinos, "Results of the Examination of Youth for Military Service, 1965," in *Supplement to Health of the Army,* July 1966, Washington.

J. D. Mooney, "Housing Segregation, Negro Employment and Metropolitan Decentralization: An Alternative Perspective," *Quart. J. Econ.,* May 1969, *83,* 299–311.

U.S. Department of Defense, (1970a) *Post-Service Earnings Among Men Who Separated from Army July 1968 through June 1969,* Manpower Res. Stud. rep. no. MA 70–2, Washington 1970; and computer printouts on which this report is based.

———, (1970b) *Unemployment Among Army Men Who Separated from Service July 1968 through June 1969,* Manpower Res. Stud. rep. no. MA 70–1, Washington 1970; and computer printouts on which this report is based.

———, *Post-Service Educational, Training, and Vocational Status of First-Term Army Men Separating During July-September 1968,* unpublished Manpower Res. Stud. rep., Washington 1969.

On Reanalyzing the Harris-Todaro Model: Policy Rankings in the Case of Sector-Specific Sticky Wages

By JAGDISH N. BHAGWATI AND T. N. SRINIVASAN*

In a brilliant and pioneering paper, John Harris and Michael Todaro introduced a model with two sectors, manufacturing (urban) and agriculture (rural), a (sticky) minimum wage in manufacturing and consequent unemployment. They also introduced a labor allocation mechanism under which, instead of the usual equalization of *actual* wages, the actual rural wage was equated with the expected urban wage; the latter was defined as the (sticky) minimum wage weighted by the rate of employment, so that, unlike in the standard rigid-wage models of trade theory (for example, Gottfried Haberler, Bhagwati, Harry Johnson, Louis Lefeber, and Richard Brecher), the unemployment resulting from the minimum wage is to be construed as *specific* to the urban sector.

In the context of this model, Harris and Todaro analyze two policies: 1) a wage subsidy policy in the manufacturing sector (alone); and 2) a labor-mobility restriction policy. They argue that the former, as well as the latter, can be used to improve welfare, defined as a function of available goods in the usual way; but that, to attain the optimal first best solution, *both* policies are necessary. The authors express regret at the necessity of using migration restrictions in view of the "... ethical issues involved in such a restriction of individual choice and the complexity and arbitrariness of administration" and end their exercise with the sentiment that:

> All of the above suggests that altering the minimum wage may avoid the problems of taxation [to finance the wage subsidy in manufacturing], administration, and interference with individual mobility attendant to the policy package just discussed. Income and wage policies designed to narrow the rural-urban wage gap have been suggested by D. P. Ghai, and Tanzania has formally adopted such a policy along with migration restriction. In the final analysis, however, the basic issue at stake is really one of political feasibility and it is not at all clear that an incomes policy is any more feasible than the alternatives. [p. 138]

We argue in this paper that this dilemma is unnecessary *in principle*, the reasons being that:

1) a *uniform* wage subsidy, regardless of the sector of employment, will yield the optimal, first best solution;

2) equivalently, a wage subsidy in manufacturing *plus* a production subsidy to agriculture will yield the optimal, first best solution;

3) in either case, no resort to "ethical compromises" in the direction of sanctioning migration restrictions will be necessary;

4) proposition 2) implies that the authors' argument that the traditional prescription to use shadow pricing of labor (i.e., a wage subsidy in employment) is inapplicable to their model is not correct and their conclusion stems from equating this prescription with the prescription that the wage subsidy be given for employment in the manufacturing sector alone; and

5) proposition 2) also implies that the authors' contention that *two* policies are necessary to attain the first best optimum is not valid *unless* one construes a general wage subsidy to constitute *two* policies when there are *two* sectors employing labor.

---

* Professors of economics, Massachusetts Institute of Technology and the Indian Statistical Institute, respectively. The paper was written while Srinivasan was visiting professor of economics at M.I.T. and the research underlying it has been supported by the National Science Foundation. The paper has profited from seminars at Minnesota, Stanford, and Harvard. Thanks are due to David McClain for computational assistance, and to Peter Diamond, John Harris, John Chipman, Anne Krueger, and a referee for helpful comments.

In demonstrating these propositions, we should note that the Harris-Todaro formal model has a demand function which is not related to the utility function in their (later) welfare analysis, so that their analytical system is open to the possibility of being overdetermined. We therefore rewrite their model, with the utility function explicitly incorporated into the model, eliminating the "additional" demand equation of Harris and Todaro.

Since the basic problems with the Harris-Todaro analysis relate to their first best optimal-policy characterization, we begin with analysis of the first best, optimal policy in the model.[1] However, we also take the opportunity to extend the analysis in Section II to two second best policy measures: wage subsidy in manufacturing and production subsidy to agriculture; both policies can be shown to be equivalent, singly or in combination, to all other conceivable policy interventions in the model. However, rather than prove these results with rigor—we have done this elsewhere in a companion paper—we produce numerical examples in the Appendix to establish and illustrate the least intuitive among them.

I. The Model

We may now restate the Harris-Todaro model. First, there are two production functions:

$$(1) \qquad X_A \leqq f_A(L_A)$$

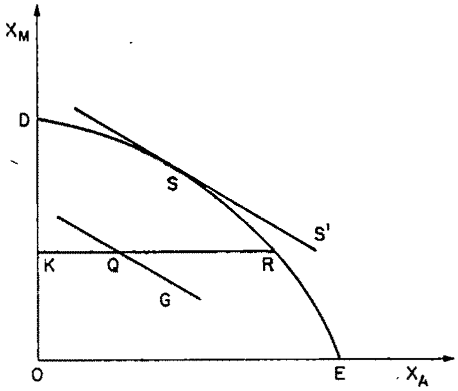
$$(2) \qquad X_M \leqq f_M(L_M)$$

where X_A and X_M are the output levels of agriculture (rural sector) and manufactures (urban sector), respectively, and L_A, L_M are the labor-input levels in the two sectors. The functions are strictly concave. The labor supply is fixed and assumed to be unity by choice of units:

$$(3) \qquad L_A + L_M \leqq 1$$

We then have a standard, social utility function:

$$(4) \qquad U = U(X_A, X_M)$$



DE is the production possibility curve when wage rigidity is absent. With the wage rigidity constraint, equilibrium production under *laissez-faire* can lie only along *RK* instead of *RD*, because equilibrium on *RD* (excluding *R*), as at *S*, implies wage in manufacturing *below* the minimum wage. *Q* is the *laissez-faire* production point under price-ratio *QG* under the wage constraint. For simplicity, the diagram depicts the price-ratio at *S* and *Q* to be identical, implying *either* a linear utility function for a closed economy *or* a "small," open economy with unchanging terms of trade. The formal analysis in the text is not restricted to linear utility functions; but it does *not* apply, without amendment, to a "large" open economy with monopoly power in trade.

FIGURE 1

where U is concave with positive marginal utilities for finite $[X_A, X_M]$.

For a fully competitive economy, the resulting equilibrium can be shown in Figure 1 at S where the production possibility curve DE is tangent to SS' and

$$(5) \qquad \frac{U_1}{U_2} = \frac{f'_M}{f'_A}$$

with U_1/U_2 equal to the negative of the slope of SS', and U_1 and U_2 representing the partial derivatives of U with respect to X_A and X_M, respectively.

But we now assume that the wage in manufacturing is fixed as a minimum, so that for this competitive economy, we must have:

$$(6) \qquad f'_M(L_M) = \bar{w}$$

If we then assume that this constraint is binding at S, the first best optimal solution is inadmissible and unemployment ensues.[2] The system could then have been characterized nonetheless by the equalization of *actual* wages in the two sectors. Harris and Todaro, however, chose to rewrite the wage-equalization equation in terms of the *expected* wage in manufacturing, defined as the actual wage there weighted by the rate of employment, so that the critical equilibrium conditions in their model, relevant for our analysis, are

---

[1] An error of detail is picked up later in this paper.

$$f'_M = \bar{w} \tag{7}$$

$$\frac{U_1}{U_2} f'_A = \bar{w} \frac{L_M}{1 - L_A} \tag{8}$$

where the total labor force is assumed to be one by choice of units and where consumption and production price of the agricultural good is identical and equal to U_1/U_2.

With $\bar{w}$ specified, (7) and (8) can be solved for L_M and L_A, using the two production functions. The *laissez-faire* equilibrium, with unemployment ($L_M < 1 - L_A$), will then lie in Figure 1 along RK (where X_M and hence L_M are fixed at the value that makes $f_M = \bar{w}$) at Q. (It may be emphasized that the *laissez-faire* equilibrium would so lie along RK even if actual wages were equalized in the two sectors: nothing critical to our interests hangs on the *expected*-wage wrinkle in the Harris-Todaro analysis.)

As for the available policy instruments (that use the price mechanism as distinct from direct allocation mechanisms) in this model, we note now the following:

(i) *laissez-faire*;
(ii) wage subsidy in manufacturing;
(iii) production subsidy to agriculture.

The structure of the model also implies the following equivalences:

(iv) a wage subsidy in agriculture is equal to policy (iii);[3]
(v) a uniform wage subsidy in all employment is a combination of policies (ii) and (iii);
(vi) for a closed economy, a consumption tax-cum-subsidy is equivalent to policy (iii), i.e., a production tax-cum-subsidy;[4]
(vii) for an open economy, a tariff (trade subsidy) policy would, as usual, be equivalent to policy (iii), i.e., a production tax-cum-subsidy policy, *plus* a consumption tax-cum-subsidy policy.[5]

One final point may be noted. Our analysis does *not* explicitly distinguish between a closed and an open economy. Since it relates essentially to the *production* equilibrium in the economy, and since it allows the utility function to be linear or non-linear, it can be interpreted as applying *either* to a closed economy *or* to an open economy with given terms of trade.[6]

---

[2] Needless to say, unemployment is inevitable only if we assume that the unemployed labor will not prefer certain employment at a lower wage in the agricultural sector to uncertain employment in the manufacturing sector at a higher wage. One has to contemplate therefore *either* a randomized process by which everyone in the manufacturing labor force gets an equal crack at the manufacturing jobs, so that *each* on the average gets the expected wage *or* that the unemployed workers return to employment in the agricultural sector. In the latter case, we could wind up with the wage-differential, full employment model which has already been extensively analyzed in trade-theoretic literature.

[3] We are assuming, in writing equation (7), that the producer and consumer prices of the manufacturing good (in terms of any arbitrary unit of account) are the same, i.e., in effect, the producers of the manufacturing good are paying the workers the wage $\bar{w}$ in kind. Hence, the effect of a production subsidy to manufacturing is essentially not to affect any real decisions, as those made *via* equations (7) and (8), but merely to increase each commodity price in terms of the (arbitrary) unit of account. However, if we were to assume instead that the producer and consumer prices of the manufacturing good could be made to differ by policy, then the worker in manufacturing would earn the value of his marginal product at the producer price and then, *qua* consumer, must have enough income (in terms of the unit of account) to buy $\bar{w}$ units of the manufacturing good. In that case, a wage subsidy policy to manufacturing would be equivalent to a production subsidy policy to manufacturing, as is the case in the agricultural sector. Thus note that, if we did shift to the latter, alternative assumption on wage payment in the manufacturing sector, then the analysis would not change but our policy equivalences would. In particular, the first best optimal policy mix would then include: a uniform production subsidy to both sectors; and a production subsidy in manufacturing and a wage subsidy in agriculture.

[4] Thus, let $\pi_p = \bar{w}L_M/(1 - L_A)f_A'$ be the production price of the agricultural good and $\pi_c = U_1[X_A, X_M]/U_2[X_A, X_M]$ be the consumption price of the agricultural good. The production tax-cum-subsidy is then $(\pi_p - \pi_c)/\pi_c$; and the consumption tax-cum-subsidy is $(\pi_c - \pi_p)/\pi_p$.

[5] Thus, if π^* is the international price of the (importable) agricultural good, a tariff at *ad valorem* rate t would imply: $\pi^*(1+t) = \pi_p = \pi_c$.

II. Optimal Policy Intervention

It is easy to see that the first best optimum can be reached in this model by the use of a uniform wage subsidy *or*, equivalently, by the use of a wage subsidy in manufacturing and a production subsidy to agriculture. Thus, let

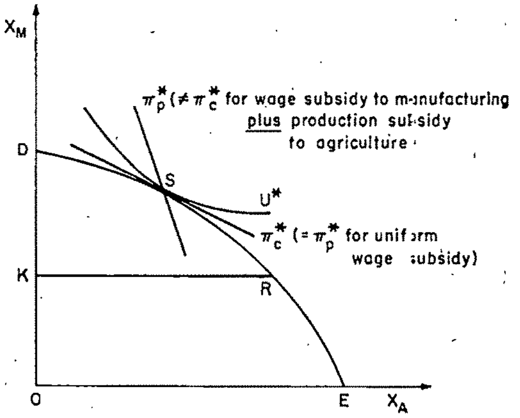
$$s^* = \bar{w} - f'_M(L^*_M)$$

be the wage subsidy (financed by appropriate lump sum taxation) in manufacturing, with the asterisks denoting first best values. If this subsidy is also extended to employment in agriculture, we should write the equilibrium condition in production as:

$$f'_M = \bar{w} - s^* \tag{9}$$

$$\pi^*_c f'_A = \bar{w} - s^* \tag{10}$$

where $\pi^*_c = U_1(X^*_A, X^*_M)/U_2(X^*_A, X_M)$ is the consumption price (equal to the producer's price $\pi^*_p = f'_M/f'_A$) of the agricultural good. It is clear then that the constraints of the model are met (i.e., the wage rate in manufacturing is at $\bar{w}$ *and* the wage rates are equalized at the producer's prices in both sectors) and full employment optimal equilibrium is reached with wage subsidy at level s^* in both sectors. Thus, in Figure 2 (which illustrates for a closed economy case), the resulting full employment, optimal equilibrium is at S, with $\pi^*_c = \pi^*_p$ (and the domestic, marginal rates of transformation in production and in consumption are equal at S).


X_M
$\pi^*_p (\neq \pi^*_c$ for wage subsidy to manufacturing plus production subsidy to agriculture
D
S
U^*
π^*_c ($= \pi^*_p$ for uniform wage subsidy)
K
R
O
E
X_A

S is the first best, optimum for a closed economy, with the social utility curve U^* tangent to the production possibility curve DE. A suitable, uniform wage subsidy to both sectors, A and M, will equate the consumption and production prices with the domestic rates of transformation in production and substitution in consumption at S. A suitable wage subsidy to manufacturing *plus* production subsidy to agriculture will not equate the consumption and production prices but will equate the two rates of substitution in consumption and transformation in production at S with each other and with the consumption price alone.

FIGURE 2

Alternatively, we could have used a wage subsidy in manufacturing (alone) at level s^* and combined it with a production subsidy in agriculture. Thus, with

$$\pi^*_p = \frac{\bar{w}}{f'_A(L^*_A)}$$

as the producer's price of the agricultural good, and π^*_c as the consumer's price of the agricultural good, as before, we have:

$$t^* = \frac{\pi^*_p - \pi^*_c}{\pi^*_c}$$

as the optimal subsidy to agriculture. With the optimal values for s^* and π^*_p, we then have:

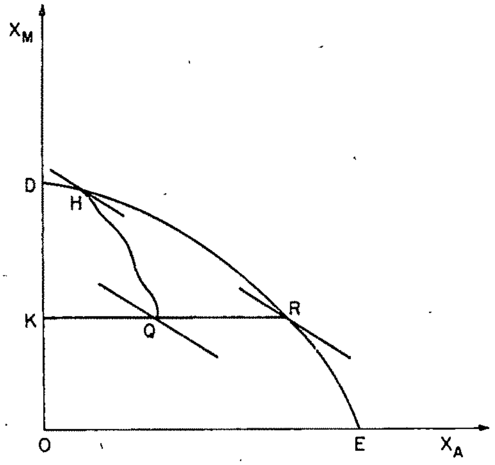
$$f'_M = \bar{w} - s^* \tag{11}$$

$$\pi^*_p f'_A = \bar{w} = \bar{w} \cdot \frac{L^*_M}{1 - L^*_A} \tag{12}$$

and, once again, we note that the constraints in the model are met, and full employment, optimal equilibrium is reached with wage

---

[6] The analysis would have to be amended to bring in the foreign reciprocal demand function explicitly into the formal model if we were to consider the case of a country with monopoly power in trade. For a "small," open economy, the analysis in the text for a linear utility function would be applicable without modification.



QH is the locus of production equilibria, traced out by increasing the wage subsidy in manufacturing from $s=0$ to $s_{\max}$ yielding full employment at *H*. *QR* is the locus of production equilibria, traced out by increasing the production subsidy to agriculture. While the diagram assumes the linear utility function case, as does Figure 1, for sake of simplicity, the formal analysis in the text is not so restricted. Note that, in the restricted case illustrated here, an increasing wage subsidy to manufacturing would necessarily reduce agricultural output; not so, in the general case which admits nonlinear utility functions.

FIGURE 3

subsidy in manufacturing at level s^* and production subsidy to agriculture at rate t^*.

However, while the equilibrium is again optimally at *S*, it is characterized now by $\pi_p^* \neq \pi_c^*$ (though of course the domestic, marginal rates of transformation in production and substitution in consumption remain equal to each other and identical to that under the uniform wage-subsidy policy at *S*).

Hence we have established the validity of criticisms 1)–5) leveled at the Harris-Todaro analysis at the outset of this paper.

III. Second Best Policy Intervention

The two second best policies which then can be considered are: a wage subsidy to manufacturing (considered by Harris-Todaro at some length); and a production subsidy to agriculture (not considered by Harris-Todaro, although their "migration-restriction" policy is the "quota-equivalent" thereof).

Wage Subsidy in Manufacturing: We only sketch here briefly the analysis of this policy as the Harris-Todaro results are totally correct.[7] With s as the wage subsidy in manufacturing, the equilibrium is now characterized by:

$$f'_M = \bar{w} - s \tag{13}$$

$$\frac{U_1}{U_2} f'_A = \bar{w} \cdot \frac{L_M}{1 - L_A} \tag{14}$$

Clearly, given $\bar{w}$ and s, (13) and (14) can be solved for L_M and L_A. We can then demonstrate (see our 1973 paper) that:

Starting from a *laissez-faire* equilibrium ($s=0$), or *RK* at *Q* in Figure 3, increasing s means shifting the production equilibrium *Q* steadily north;

the locus of successive production equilibria, mapped out by increasing s, must reach full employment (at an $s_{\max}$) on the production possibility curve: such a locus being *QH*;[8]

the full employment equilibrium may be inferior welfarewise to *laissez-faire*—a proposition which we illustrate with a numerical example in the Appendix;

a wage subsidy will necessarily improve welfare (i.e., $dU/ds > 0$ at $s=0$); and

the second best wage subsidy need not be characterized by full employment, so that tradeoff possibilities between increased welfare (*via* a standard social utility function of the type deployed by Harris and Todaro, and in this paper) and reduced unemployment may be pertinent.

Production Subsidy to Agriculture: For the case where the policy instrument is a production subsidy to agriculture, the equilibrium conditions are clearly rewritten as:

---

[7] We have also developed the second best analysis at much greater length, and with formal rigor, in the companion paper cited earlier. Instead, we give numerical examples in the Appendix to illustrate the major propositions listed here.

[8] Harris and Todaro incorrectly argue, p. 134, that the full employment equilibrium with a wage subsidy in manufacturing can be inside *DE*, off the production possibility curve. They forget that labor is the only factor in the model, in effect; they seem to have erred by relying on analogy with the standard two-factor model.

$$f'_M = \bar{w} \tag{15}$$

$$\pi_p f'_A = \frac{\bar{w} L_M}{1 - L_A} \tag{16}$$

where, as before, π_p is the producer's price of the agricultural good and the implied production subsidy is $(\pi_p - U_1/U_2)/(U_1/U_2)$. Clearly, given π_p and $\bar{w}$, we can solve for L_M and L_A. It is also then easy to show that:

Starting from a *laissez-faire* equilibrium ($\pi_p = U_1/U_2$), on RK at Q in Figure 3, increasing π_p will steadily shift the production equilibrium to the right along QR until full employment is reached at $\bar{\pi}_p$ at R;

the equilibrium at R is also the second best optimal equilibrium, so that the full employment, second best equilibrium is reached when $\pi_p = \bar{\pi}_p$ and there is an implied production subsidy to agriculture; and

the second best wage subsidy in manufacturing cannot be ranked uniquely with the second best production subsidy to agriculture—as illustrated by a numerical example in the Appendix.

IV. Concluding Remarks

Where do the migration-restriction policies of Harris and Todaro fit in?

If one is willing to contemplate direct, physical allocations, one can clearly reach the first best, optimal solution, S in Figure 1, by assigning the corresponding labor to the two sectors (L_A^* and L_M^*) and enforcing the rule that all labor be employed regardless of private profitability (thus yielding X_A^* and X_M^*). The Harris-Todaro policy package for reaching S, consisting of a wage subsidy in manufacturing and migration restrictions is thus a "mixed" package: one policy being of the price-mechanism variety and the other of the direct physical-mechanism variety. One could equally turn this mixed package on its head and have manufacturers forced to employ all available labor and let a production subsidy to agriculture allocate the labor force at the optimal values (L_A^* and L_M^*).

Nothing can be said, in principle, about the relative suitability of all these equivalent alternatives without bringing in other considerations, including the ethical considerations mentioned by Harris and Todaro, to introduce asymmetries/nonequivalences among them.

Further, as for second best policies, we might be able to justify the Harris-Todaro concentration on the wage subsidy to manufacturing policy, as against a uniform wage subsidy policy, on feasibility grounds. It may well be that the government's capacity to intervene is confined to the (modern) urban sector and that a wage subsidy in agricultural employment is infeasible. This is, however, a question of empirical import, and it does not really justify the exclusion from the theoretical analysis of the first best price-mechanism variety intervention.

Finally, we may note explicitly that an attempted extension of our policy rankings to actual policy implementation would have to take into account the following, well-known problems:

The administration costs *and* feasibility of alternative policies must be taken into account.

Since taxes must be collected to disburse subsidies, the question arises whether those who ask for minimum real wages will not, even when such taxes are imposed on them in a lump sum fashion, seek to revise the minimum real wage that is demanded. We have assumed, of course, that the minimum real wage demanded is independent of the tax policy chosen.

APPENDIX

We produce numerical examples to show that:

A full employment yielding wage subsidy in manufacturing may be inferior to *laissez-faire*.

The second best wage subsidy in manufacturing may be inferior or superior to the second best production subsidy to agriculture; the two policies cannot be uniquely ranked.

Let us consider the following production and utility functions: $f_A(L_A) = L_A^{0.75}$, $f_M(L_M) = L_M^{1/2}$, $U = pX_A + X_M$. Let p take two al-

TABLE 1

		$p=1.5$	$p=0.5$
Minimum Wage $=(L_M^*)^{-1/2}$		2.363709	1.119195
First Best Optimum	L_A	0.821017	0.201660
	X_A	0.862510	0.300929
	L_M	0.178983	0.798340
	X_M	0.423064	0.893499
	U	1.716828	1.043963
Laissez-Faire Equilibrium	L_A	0.908222	0.499286
	X_A	0.930345	0.593967
	L_M	0.044746	0.199585
	X_M	0.211532	0.446749
	U	1.607048	0.743733
Second Best Wage Subsidy Equilibrium	L_A	0.904517	0.012604
	X_A	0.927497	0.037617
	L_M	0.046600	0.987396
	X_M	0.215869	0.993678
	U	1.607114	1.012486
Full Employment Wage Subsidy Equilibrium	L_A	0.051314	0.012604
	X_A	0.107814	0.037617
	L_M	0.948684	0.987396
	X_M	0.974004	0.993678
	U	1.135726	1.012486
Second Best Production Subsidy Equilibrium	L_A	0.955254	0.800415
	X_A	0.966249	0.846226
	L_M	0.044746	0.199585
	X_M	0.211532	0.446749
	U	1.660906	0.869862

ternative values 1.5 and 0.5. Let the specified minimum wage (in terms of manufactured good) in manufacturing be twice the equilibrium wage associated with the first best optimum. The following table gives the equilibrium factor allocations, output, and welfare associated with each of the following policies: first best optimum; *laissez-faire*; second best wage subsidy to manufacturing; full employment wage subsidy to manufacturing; and second best production subsidy to agriculture.

It is seen that when $p=0.5$, the second best optimum wage subsidy[9] to manufacturing happens to be the full employment wage subsidy, and it dominates the second best production subsidy (to agriculture) whereas, when $p=1.5$, the second best production subsidy dominates the second best wage subsidy. Further, the full employment wage subsidy is inferior to *laissez-faire*, when $p=1.5$.

---

[9] Note that the values for U in the table represent a global maximum because p is fixed.

REFERENCES

J. Bhagwati, "The Theory and Practice of Commercial Policy," in *Special Papers in International Economics*, no. 8, Princeton 1968, 1–69.

——— **and T. N. Srinivasan**, "The Ranking of Policy Interventions Under Factor Market Distortions: The Case of Sector-Specific Sticky Wages and Unemployment," *Sankhya* 1973, forthcoming.

R. Brecher, "Minimum Wage Rates and the Theory of International Trade" unpublished doctoral dissertation, Harvard Univ. 1971.

G. Haberler, "Some Problems in the Pure Theory of International Trade," *Econ. J.*, June 1950, *60*, 223–40.

J. Harris and M. Todaro, "Migration, Unemployment and Development: A Two-Sector Analysis," *Amer. Econ. Rev.*, Mar. 1970, *60*, 126–42.

H. G. Johnson, "Optimal Trade Intervention in the Presence of Domestic Distortions," in R. E. Caves et al., eds., *Trade, Growth, and the Balance of Payments*, Amsterdam 1965, 3–34.

L. Lefeber, "Trade and Minimum Wage Rates," in J. Bhagwati et al., eds., *Trade, Balance of Payments, and Growth*, Amsterdam 1971, 91–114.

The Nature of Quantities in Market Disequilibrium

By HERSCHEL I. GROSSMAN*

The Walrasian paradigm of tâtonnement with recontracting depicts a frictionless system of markets in which potential buyers and sellers have full information about all relevant alternatives and exchange takes place only at prices consistent with general market clearing. Much recent theoretical research has been concerned with modifying this elegant, but heroically abstract, paradigm to allow for various forms of friction. As one aspect of this wide-ranging endeavor, several authors have discussed the interaction of price and "quantity" adjustments in a single market in which transactions generally take place under non-market-clearing conditions. Examples of such discussions include the seminal paper by Martin Beckmann and Harl Ryder and the more recent paper by E. C. H. Veendorp.

The present note is concerned with the inadequacy of these contributions. It develops the following theme: In any disequilibrium situation, quantity is a multifaceted concept and the determination of quantity is a multifaceted process. Unfortunately, previous discussions, including the contributions mentioned above, have failed to appreciate the complexity of this problem. The consequent ambiguity regarding the concept of quantity has resulted in ambiguous and essentially uninteresting analysis as well as confused conclusions. Clear distinctions between the various relevant concepts of quantity are a precondition for useful analysis of the substantive issues involved in disequilibrium dynamics. The essential questions are both substantive and semantic.

Economic analysis deals with a number of different quantities, which, in disequilibrium situations, and especially in non-market-clearing disequilibria, need not be equivalent. What are these different quantities? Of proximate interest on the buying side, we may distinguish, at least, between 1) the quantity demanded; 2) the quantity ordered; 3) the quantity bought and delivered; and 4) the quantity consumed. Of proximate interest on the selling side, we may distinguish, at least, between 1) the quantity supplied; 2) the quantity offered; 3) the quantity sold and delivered; and 4) the quantity produced. In what follows, Section I discusses the meanings of these concepts, the distinctions between them, and processes suggested by economic theory for quantitatively determining them. Section II refers to the contributions of Beckmann and Ryder and Veendorp to illustrate the sort of confusion which results from the failure to distinguish clearly between the relevant concepts of quantity.

I. The Concepts of Quantity

Consider a market for a homogeneous good or service. Assume that a Walrasian-type auctioneer proximately determines price, so that both buyers and sellers are price takers.[1] The four quantity concepts relevant to each of the buying and selling sides of the market are symmetrical. Consider each of the four pairs in turn.

1) The quantities demanded and supplied: The quantity demanded is the quantity which if bought would be consistent with the maximization of buyer objectives, given the existing price and any other constraints on buyer behavior. These other constraints involve technology, predetermined endowments, and price and quantity data generated in other markets.[2] If exchange is volun-

---

* Professor of economics, Brown University. The National Science Foundation supported this research. I have benefited from the comments of John Black.

[1] Alternative frameworks in which quality is not uniform or in which either buyers or sellers are price setters introduce other important and difficult considerations, which for present purposes are best avoided.

[2] We are not concerned here with explicitly specifying the interrelationships between markets which may be failing to clear. In this regard, see, for example, the author (1969, 1971).

tary, the quantity demanded is also the maximum quantity which buyers would buy at the existing price.[3]

Quantity supplied is analogous to quantity demanded. The quantity supplied is the quantity which if sold would be consistent with the maximization of seller objectives, given the existing price and any other constraints on seller behavior. If exchange is voluntary, the quantity supplied is also the maximum quantity which sellers would sell at the existing price.

If the adjustment of either the quantity demanded or the quantity supplied is costly—say, because decision making is unpleasant—long-run and short-run demand and supply may differ. Simple linear specifications of the relationships which determine quantities demanded and supplied might be the following:[4]

$$(1)\quad D^* = \alpha_D + \beta_D P, \qquad \alpha_D > 0, \quad \beta_D < 0$$

$$(2)\quad \dot{D} = \lambda_D(D^* - D), \qquad \lambda_D > 0$$

$$(3)\quad S^* = \alpha_S + \beta_S P, \qquad \alpha_S < 0, \quad \beta_S > 0$$

$$(4)\quad \dot{S} = \lambda_S(S^* - S), \qquad \lambda_S > 0$$

where P is price, D^* is long-run demand, D is short-run demand, S^* is long-run supply, and S is short-run supply.

2) The quantities ordered and offered: The quantity ordered measures the buyers' active attempts to purchase. If buyers are able to buy the quantity demanded, we would expect the quantity ordered to be equal to the quantity demanded. However, if the aggregate quantity demanded exceeds the quantity supplied and exchange is voluntary so that the quantity bought cannot exceed the quantity supplied, on average, buyers are not able to buy the quantity demanded. In that event, they are likely to order a quantity either greater or less than the quantity demanded.[5] As one possibility, if individual buyers think that quantities actually bought and delivered are directly related to the quantity ordered, they may order a quantity greater than the quantity demanded, in order to increase their chances of actually buying the quantity demanded.[6] Alternatively, if there are costs associated with the placing of orders, buyers may not bother to place orders which are expected to be unsuccessful. In this case the quantity ordered would be less than the quantity demanded.

The quantity offered measures the sellers' active attempts to sell. The analysis of the quantity offered is analogous to the analysis of the quantity ordered. If sellers are able to sell the quantity supplied, we would expect them to offer the quantity supplied. However, if the aggregate quantity supplied exceeds the quantity demanded, and exchange is voluntary so that the quantity sold cannot exceed the quantity demanded, the quantity offered may be either greater than or less than the quantity supplied.[7]

Simple linear specifications of the relationships which determine quantities ordered and offered might be the following:[8]

---

[3] This is the standard textbook definition of demand. See, for example, Milton Friedman, p. 13.

[4] This note contains no formal derivations of quantity-determining processes. In some cases, these derivations are trivial. In other cases, they are well known or, at least, treated elsewhere. For example, the articles by Zvi Griliches and Robert Barro discuss models which can rationalize the form of the adjustment processes suggested below.

[5] The present discussion abstracts from the process by which buyers and sellers learn whether they are or are not able to buy and sell the quantities demanded and supplied. Presumably, buyers and sellers readily ascertain the existence of a non-market-clearing situation, if this situation is not transitory.

[6] Bent Hansen's important study of inflation emphasized this possibility.

[7] The distinction between quantity supplied and quantity offered is particularly relevant to evaluating the measured labor force and unemployment statistics reported by the U.S. Bureau of Labor Statistics. The *BLS* survey procedure defines the labor force as the quantity offered. However, the actual constraint on employment is the quantity supplied. The measured labor force might exceed the quantity supplied because, for example, unemployment of primary workers induces secondary workers to enter the market. The quantity supplied might exceed the measured labor force because of the discouraged worker effect.

[8] As with quantities demanded and supplied, if adjustment is costly, long-run and short-run orders and offers could differ. Allowing for this distinction would complicate the analysis considerably. For example, equation (8) below would no longer be generally appropriate.

$$(5)\quad \begin{cases} d = D, & \text{if } D \leqq S \text{ and} \\ d = D + \beta_d(D - S), & -1 < \beta_d \gtreqless 0, \\ & \text{if } D > S \end{cases}$$

$$(6)\quad \begin{cases} s = S, & \text{if } D > S \text{ and} \\ s = S + \beta_s(D - S), & 1 > \beta_s \gtreqless 0, \\ & \text{if } D < S \end{cases}$$

where d represents orders and s represents offers.

Special interest attaches to the quantities d and s because they represent the only information regarding buyer and seller intentions which is naturally and generally available to the auctioneer responsible for setting the price. The auctioneer necessarily observes the quantity D only if $D \leq S$, so that $d = D$. The auctioneer necessarily observes the quantity S only if $D \geq S$, so that $s = S$. Thus, in general, the auctioneer does not observe the quantities D and S. Assuming that the auctioneer is searching for a market-clearing price—that is, a price which is consistent with equality between d and s, and thus which is also consistent with equality between D and S—and assuming that the auctioneer believes that in general the difference d minus s is inversely related to P, we suppose that he adjusts P according to

$$(7)\qquad \dot{P} = \lambda_P(d - s), \qquad \lambda_P > 0$$

3) The quantities bought and sold: The quantities bought, sold, and delivered must, of course, be identical and may be referred to as the quantity transacted. If exchange is voluntary, the quantity transacted cannot be greater than the minimum of the quantities demanded and supplied—that is, $X \leq \min(D, S)$, where X is the quantity transacted.[9] If exchange exhausts all mutually advantageous transactions, then the quantity transacted cannot be less than the minimum of the quantities demanded and supplied—that is, $X \geq \min(D, S)$. Combining the assumptions of voluntary and exhaustive exchange yields[10]

$$(8)\qquad X = \min(D, S)$$

4) The quantities produced and consumed: If the commodity in question is not storable, then the quantity consumed cannot exceed the quantity transacted and the quantity transacted cannot exceed the quantity produced—that is, $C \leq X \leq Y$, where C is the quantity consumed and Y is the quantity produced. Moreover, because wastage is nonoptimal for both buyers and sellers, production would never exceed sales and purchases would never exceed consumption—that is, $C \geq X \geq Y$. Consequently,

$$(9)\qquad C = X$$

$$(10)\qquad Y = X$$

Alternatively, if the good in question is storable, neither consumption nor production need be equal to the quantity transacted. If adjustment of either the quantity produced or consumed is costly, long-run consumption and production both of which are presumably equal to the quantity transacted may differ from short-run consumption and production, these differences being absorbed by inventories. Simple linear specifications of the adjustment of consumption and production might be the following:

$$(11)\qquad \dot{C} = \lambda_C(X - C), \qquad \lambda_C > 0$$

$$(12)\qquad \dot{Y} = \lambda_Y(X - Y), \qquad \lambda_Y > 0$$

Equations (1) through (12), taken together, represent a consistent theory of the adjustment over time of price and the various relevant quantities in a single competitive market. The complete model consists of two first-order linear systems, one

---

[9] Examples of markets in which, historically, exchange has not been strictly voluntary include the market for military manpower and the markets for the services of certain regulated public utilities. This concept of voluntary exchange implies that orders and offers are not binding. If orders and offers were binding, we would have $X \leq \min(d, s)$. This discussion also ignores the complication of long-term contracts to buy or sell, which, in the event of unpredicted shifts in constraints, buyers or sellers might regret having entered into.

[10] If the exhaustion of mutually advantageous transactions is a gradual process, we would have to replace equation (8) with the following discontinuous adjustment relationship: $X \leq \min(D, S)$ and $\dot{X} = \lambda_x[\min(D, S) - X]$.

$$(13)\qquad \begin{bmatrix} \dot{P} \\ \dot{D} \\ \dot{S} \\ \dot{C} \\ \dot{Y} \end{bmatrix} = \begin{bmatrix} 0 & \lambda_P(1-\beta_s) & -\lambda_P(1-\beta_s) & 0 & 0 & 0 \\ \lambda_D\beta_D & -\lambda_D & 0 & 0 & 0 & \lambda_D\alpha_D \\ \lambda_S\beta_S & 0 & -\lambda_S & 0 & 0 & \lambda_S\alpha_S \\ 0 & \lambda_C & 0 & -\lambda_C & 0 & 0 \\ 0 & \lambda_Y & 0 & 0 & -\lambda_Y & 0 \end{bmatrix} \begin{bmatrix} P \\ D \\ S \\ C \\ Y \\ 1 \end{bmatrix}$$

$$(14)\qquad \begin{bmatrix} \dot{P} \\ \dot{D} \\ \dot{S} \\ \dot{C} \\ \dot{Y} \end{bmatrix} = \begin{bmatrix} 0 & \lambda_P(1+\beta_d) & -\lambda_P(1+\beta_d) & 0 & 0 & 0 \\ \lambda_D\beta_D & -\lambda_D & 0 & 0 & 0 & \lambda_D\alpha_D \\ \lambda_S\beta_S & 0 & -\lambda_S & 0 & 0 & \lambda_S\alpha_S \\ 0 & 0 & \lambda_C & -\lambda_C & 0 & 0 \\ 0 & 0 & \lambda_Y & 0 & -\lambda_Y & 0 \end{bmatrix} \begin{bmatrix} P \\ D \\ S \\ C \\ Y \\ 1 \end{bmatrix}$$

which describes price and quantity adjustments when P is above the market-clearing price and one which describes price and quantity adjustments when P is below the market-clearing price. The full dynamic system is summarized by equation system (13) for $X=d=D<S\neq s$, and by equation system (14) for $d\neq D>S=s=X$. Note also that since D and S are generally changing over time, the boundary between the domains of equations (13) and (14) will also be changing over time.

Clearly this model allows for a wide variety of patterns of motion. An analysis of the range of possibilities is beyond the scope of the present note.[11] The purpose of the present discussion is to clarify the relevant concepts and to facilitate the identification of the assumptions implicit in more simplified formulations of price and quantity adjustment. To illustrate the importance of this clarification, I now turn to a specific evaluation of previous contributions.

II. Previous Discussions of Price and Quantity Adjustments

Beckmann and Ryder present two models of simultaneous price and quantity adjustment. They regard those models as formally equivalent. Their first model deals with the product market, and is specified as

$$(15)\qquad \dot{P} = \lambda[f(P) - Q]$$

$$(16)\qquad \dot{Q} = \mu[P - q(Q)]$$

where P is price, Q is quantity, $f(P)$ is quantity demanded, and $g(Q)$ is marginal cost. The essential problem involved in interpreting this model is the ambiguity of the variable Q. The question at hand is the following: Do equations (15) and (16) represent a logically consistent model? More specifically, what interpretation if any may be given to the variable Q in order to make equations (15) and (16) consistent with the model summarized by equations (13) and (14)?

The answer is that equations (15) and (16) do represent a logically consistent model and a specific simplification of equations (13) and (14) if Q is interpreted as quantity supplied. The nature of this simplification is readily apparent: First, abstract from the distinctions between quantity ordered and quantity demanded and between quantity offered and quantity supplied, so that equations (5) and (6) simplify to $d=D$ and $s=S$. Second, assume that λ_D is very large, so that equation (2) simplifies to $D=D^*$. Then substituting these relationships into equation (7) yields

---

[11] Beckmann and Ryder and Veendorp illustrate the application of the mathematical techniques relevant to such an analysis.

$$\dot{P} = \lambda_P(D^* - S)$$

which is equivalent to equation (15), where $f(P) \equiv D^*$, $Q \equiv S$, and $\lambda \equiv \lambda_P$. Finally, observe that equation (16) represents an alternative linear approximation to equation (4), because $S^* - S$ and $P - g(Q)$ have the same sign, where $Q \equiv S$ and S^* represents the level of sales consistent with equality between P and $g(Q)$.

However, although equations (15) and (16) represent a logically consistent system given that Q is interpreted as quantity supplied, they are seriously deficient as a descriptive model of non-market-clearing situations. Specifically, equations (15) and (16) themselves tell us nothing about quantities produced, sold, and consumed.[12] An explanation of these quantities requires, in addition to a theory of quantities demanded and supplied, a theory of transactions exemplified by equation (8), and theories of inventory accumulation and decumulation exemplified either by equations (9) and (10) or by equations (11) and (12).

The second model presented by Beckmann and Ryder deals with the factor market, and is specified as

$$\dot{P} = \lambda[q - h(P)] \tag{17}$$

$$\dot{q} = \mu[j(q) - P] \tag{18}$$

where P is the wage rate, q is quantity, $h(P)$ is quantity supplied, and $j(Q)$ is marginal revenue product. Again, this model is logically consistent and a specific simplification of equations (13) and (14), but here q must be interpreted as quantity demanded. Assuming that $d = D$, that $s = S$ and that λ_S is very large, equation (17) is equivalent to equation (7), where $q \equiv D$, $h(P) \equiv S^*$, and $\lambda \equiv \lambda_P$. Moreover, equation (18) corresponds to equation (2), where $q \equiv D$ and D^* represents the level of employment consistent with $j(q) = P$.[13]

Analogously to equations (15) and (16), equations (17) and (18), although logically consistent, tell us nothing about the level of transactions—which, in this case, corresponds to the level of employment. Moreover, the formal equivalence between their two models, claimed by Beckmann and Ryder, seems to be superficial, because each model implies a different interpretation of the concept of quantity.

The point of departure of Veendorp's discussion is the first model proposed by Beckmann and Ryder. Veendorp objects that the formulation represented in equation (16) is not a satisfactory representation in an excess supply situation.

> . . . market supply may exceed demand in which case sellers are unable to sell the desired amount. But these sellers will, according to equation [(16)], continue to increase their supply as long as price exceeds marginal cost, no matter what percentage of their supply can be sold. [p. 1011]

As an alternative to equation (16) for excess supply situations, Veendorp proposes

$$\dot{Q} = \mu[Pf(P)/Q - g(Q)], \quad \text{for } f(P) < Q \tag{19}$$

Here, P is replaced by the product of P and the ratio $f(P)/Q$.

Taken at face value, Veendorp's criticism of equation (16) is simply wrong. Quantity supplied, which at the existing price is both the profit maximizing quantity and the maximum quantity which the firm would sell, should indeed depend only upon the existing price and upon cost considerations. The maximization calculus which determines the quantity supplied is completely independent of the quantities demanded and transacted. In contrast, the quantity offered or the quantity produced may well depend on demand considerations. Interestingly, although Beckmann and Ryder refer to Q ambiguously as "quantity," Veendorp specifically refers to Q as "quantity supplied."

---

[12] A casual reader of the paper by Beckmann and Ryder might be tempted to interpret Q as the quantity produced and to interpret equation (16) as an adjustment relationship between quantity produced and quantity supplied. However, such an interpretation in general would not be logically consistent, because equality between quantity produced and quantity supplied is rational only if price is not above the market-clearing level.

[13] Daniel Hamermesh analyzes essentially the same model, but he is clearer in specifying the relevant quantity to be the quantity demanded.

What more can be said about equation (19)? In discussing this equation, Veendorp assumes that, in an excess supply situation, quantity sold equals quantity demanded, which seems correct, while quantity produced equals quantity supplied. As he states, ". . . only $100r[f(P)/Q]$ percent of their supply can be sold, whereas the remainder disappears by spoilage . . ." (p. 1012). Perhaps, one could rationalize equation (19) by interpreting Q as the quantity produced, under Veendorp's restriction that quantity produced equals quantity supplied. However, this restriction is hardly plausible. Why should the quantity produced be equated to the quantity supplied, when the quantity supplied exceeds quantity sold? In any event, Veendorp's discussion, including his criticism of equation (16) and his formulation of equation (19), seems essentially to reflect a confusion between the concepts of quantity supplied and quantity produced.

In summary, the discussions by Beckmann and Ryder and Veendorp exemplify the ambiguity and confusion which follows from failure to appreciate the multifaceted nature of the concept of quantity which is inherent in a non-market-clearing context. With appropriate clarification, the Beckmann-Ryder models are logically consistent, but less informative than they superficially appear. Veendorp's analysis, however, seems to be logically inconsistent. Even in a single market context, a model of the interaction of price and quantity adjustments, which is both logically consistent and empirically interesting, requires the specification of an extensive set of quantity concepts and corresponding behavioral relationships. These relationships, as exemplified by the model summarized in equations (13) and (14), must describe the activities of buyers and sellers, the behavior of the price-setting agent, and the transactions rules governing the market.

REFERENCES

R. Barro, "A Theory of Monopolistic Price Adjustment," *Rev. Econ. Stud.*, Jan. 1972, *39*, 17–26.

M. J. Beckmann and H. E. Ryder, "Simultaneous Price and Quantity Adjustment in a Single Market," *Econometrica*, July 1969, *37*, 470–84.

M. Friedman, *Price Theory*, Chicago 1962.

Z. Griliches, "Distributed Lags: A Survey," *Econometrica*, Jan. 1967, *35*, 16–49.

H. I. Grossman, "Theories of Markets Without Recontracting," *J. Econ. Theory*, Dec. 1969, *1*, 476–79.

———, "Money, Interest, and Prices in Market Disequilibrium," *J. Polit. Econ.*, Sept./Oct. 1971, *79*, 943–61.

D. S. Hamermesh, "Price and Quantity Adjustment in Factor Markets," *Western Econ. J.*, Mar. 1973, *11*, 118–25.

B. Hansen, *A Study in the Theory of Inflation*, New York 1951.

E. C. H. Veendorp, "Price Quantity Adjustments in a Competitive Market," *Amer. Econ. Rev.*, Dec. 1972, *62*, 1011–15.

The Nature of Quantities in Market Disequilibrium

By E. C. H. VEENDORP*

In his interesting paper on "The Nature of Quantities in Market Disequilibrium" Herschel Grossman correctly observes that in disequilibrium situations "quantity" is a multifaceted concept, and that the determination of quantity is a multifaceted process. Having labeled and discussed these facets, he formulates a price-quantity adjustment mechanism which is then used to evaluate the mechanisms proposed by Martin Beckmann and Harl Ryder and by the author. As far as the latter is concerned, Grossman concludes 1) that my criticism of the Beckmann-Ryder model is wrong and that my modification of their model is ambiguous, because my quantity concepts do not fit into the classification he proposes; and 2) that my adjustment process is logically inconsistent since it can not be derived as a special case of his own. I shall discuss these two charges in this order.

I

On the supply side of the market—the one relevant for the present discussion—Grossman distinguishes between the quantity supplied (S), the quantity offered (s), the quantity sold (X), and the quantity produced (Y). The label "quantity supplied" should, according to Grossman, be reserved for the quantity that sellers would like to supply at a given price on the assumption that all units can be sold, since such usage corresponds to text book terminology. It should be noted, however, that these textbook discussions are essentially limited to equilibrium situations; and to the extent that disequilibrium is discussed adjustments are usually of the tâtonnement variety. Under these conditions the different quantity concepts tend to merge, and, specifically, no distinction remains between quantity supplied and quantity offered.

I fail to see therefore, why the term quantity supplied (and/or "supply") should be used exclusively for one particular facet when the analysis progresses to disequilibrium economics, especially when the relevance of that facet is considerably reduced and has, therefore, in a different but closely related context been referred to as "notional supply." My use of the term seems at least equally defensible: given the fact that both of our papers deal with an auctioneer-type market, an institution which is usually relied upon mainly for its informational convenience, it seems quite plausible and natural to use "quantity supplied," "quantity demanded," and "price" for the signals that are sent back and forth between traders and auctioneer.

But even if one accepts Grossman's terminology as the only appropriate one, I doubt whether my "misuse" of his labels is likely to have confused the more than casual reader of my paper. My contribution is basically an extension of the Beckmann-Ryder paper, and the reader is therefore justified in expecting a similar terminology. I do believe that the supply variable used by these authors represents quantity produced and offered for sale. This interpretation is supported by their reference to a "Marshallian" process,[1] and by the more detailed discussion of the adjustment mechanism in a closely related paper by Beckmann and James Wallace.[2]

Taking my paper by itself, it should have been equally clear that my quantity variable was meant to represent quantity produced and offered for sale. Any other interpretation would be inconsistent with (a) my formulation of the price mechanism, (b) the lack of any other explanation of the asymmetry between buyer and seller response which is typical for my model (and the one of Beckmann and Ryder),[3] and (c) the fact that I explicitly limited my analysis to perishable commodities. While I am willing to admit that I could have been more explicit in defining my variables, I do not think that my paper leaves much room for terminological confusion.

---

* Tulane University. I wish to thank Herschel Grossman for his comments on an earlier version of this paper.

[1] See Alfred Marshall, book 5, ch. 5, pp. 363–80.

[2] This adjustment mechanism explicitly permits unintentional building up of inventories when quantity supplied exceeds current demand. See Beckmann and Wallace, p. 941.

II

Grossman's second point of criticism is that my model is logically inconsistent since it can not be obtained as a simplified case of his own. I believe, however, that such consistency is not crucial or even desirable because his model contains, in my opinion, a number of peculiar elements and is certainly less general than he tends to suggest.

1) As argued above, the quantity supplied as defined by Grossman has little relevance for those decision-making units that are affected by market disequilibrium. Grossman assumes that these units first maximize their objective functions under the hypothetical condition of market clearing, and then convert these optimal solutions to market signals by a mechanical adjustment that allows for disequilibrium. I would prefer a formulation according to which these signals are directly determined by some maximizing process using all relevant information concerning market conditions (prices and realization rates), and in which Grossman's demand and supply concepts would play no independent role.

2) Traditionally an auctioneer makes his pricing decisions on the basis of quantities ordered and offered by individual traders, who in their turn use these market prices to adjust their demand and supply signals. In the Grossman model, however, traders are also informed about the (notional) market excess demand, $D-S$, only one component of which represents a traditional market signal.[4] I find it difficult to visualize a process by which such information is transmitted among traders, and even more difficult to understand why none of these messages can be intercepted by the auctioneer. This functionary is a rather pitiful and bloodless creature as things stand, motivated not by any maximizing behavior, but by a doglike sense of duty; Grossman seems eager to degrade him to a complete puppet whose actions can hardly any longer be taken as an approximation of market forces. If one is willing to retain the auctioneer for transmitting price information, then it seems natural, and certainly informationally efficient, to make him the source of other market information.

3) To make things worse (for the auctioneer) the market signals are nonbinding. To determine the quantity sold, Grossman invokes the principle of voluntary exchange which he formulates as $X \leq \min (D, S)$, but the alternative $X \leq \min (d, s)$ (to which Grossman refers in a footnote) may be preferable: if the signals given to the auctioneer are not binding, then the process could easily degenerate to a game of "pick the biggest (or smallest) number."

4) Grossman does not permit wastage and, for the case of perishable commodities, equates quantity sold (X) and quantity produced (Y). This assumption implies that all production is on order, which makes the analysis somewhat less than perfectly general. And since my model explicitly deals with a perishable commodity that has to be produced before it can be sold, its logical inconsistency in the sense of Grossman is hardly surprising.

Taking into account these four observations one could easily formulate alternatives to the Grossman model.[5] If so desired, my adjustment mechanism could be derived as a special case of such an alternative version making simplifications similar to those by which Grossman tries to produce the Beckmann-Ryder model as a special case of his own.

---

[3] If the asymmetry applied to notional demand and supply, one would have to explain why it is more "unpleasant" (see Grossman, p. 510) to make supply than demand decisions.

[4] If, for example, $S>D$ then $D=d$ but $S=\alpha_s S+\beta_s (D-S))$. This would seem to suggest that only sellers need supplementary information concerning notional supplies of other sellers. However, even buyers do not know what to signal to the auctioneer until they are informed about the sign of $D-S$.

[5] The observations specifically suggest to replace Grossman's equations (1)–(6) by $d^*=f(P, d, s)$, $s^*=g(P, d, s)$, $\dot{d}=\lambda_d(d^*-d)$, and $\dot{s}=\lambda_s(s^*-s)$, and to replace (8) by $X=\min (d, s)$.

REFERENCES

M. J. Beckmann and H. E. Ryder, "Simultaneous Price and Quantity Adjustment in a Single Market," *Econometrica,* July 1969, *37,* 470–84.

——— and J. P. Wallace, "Marshallian Versus Walrasian Stability," *Kyklos,* fasc. 4, 1967, *20,* 935–48.

H. I. Grossman, "The Nature of Quantities in Market Disequilibrium," *Amer. Econ. Rev.,* June 1974, *64,* 509–14.

A. Marshall, *Principles of Economics,* 8th ed., London 1920.

E. C. H. Veendorp, "Price-Quantity Adjustments in a Competitive Market," *Amer. Econ. Rev.,* Dec. 1972, *62,* 1011–15.

IN MEMORIAM

LUDWIG VON MISES

1881–1973

Ludwig von Mises died on October 10, 1973, in New York City at the age of 92. His career as a writer on economics began in 1902 with a monograph in German on *The Development of the Feudal-Peasant Relationship in Galicia: 1772–1848* and continued, with the best-known books appearing in 1912, 1922, and 1949, for almost seventy years. His career as a teacher of economics began officially in 1913, the year before World War I diverted him to a military career of four years, but actually only after the end of the war, in 1918, at the University of Vienna; it continued after 1933 for seven years at the Institut Universitaire de Hautes Etudes Internationales in Geneva, and it ended only in 1969 at the New York University, when Mises was 88 years old.

Mises was certainly not a popular economist; by his blunt criticism of popular views and policies, by his unrelenting attacks on inflationism, interventionism, and socialism, and by his uncompromising steadfastness in arguing the case for private enterprise and free markets, he acquired as many intellectual enemies and detractors as any of the renowned economists of the twentieth century. At the same time, Mises was a beloved teacher and friend of a host of students who came to appreciate the integrity and profundity of his teaching in courses and seminars but particularly in his private seminars. The famous "Mises Circle" in Vienna recruited or produced a remarkable number of respected scholars in philosophy, politics, and social theory, as well as economics.

Of the sixteen books by Mises (or twenty-seven including revised editions), three are regarded as milestones in the history of economic thought. *The Theory of Money and Credit* appeared in its first German edition in 1912 (first English edition in 1934). To appreciate the importance of this book one must compare it with the literature of the times. In contrast to the legalistic treatment of money in the German literature of the day and to the mechanistic treatment in much of the American writings on money—the former neglecting all effects of monetary policy upon prices and economic activity, the latter focusing chiefly on the effects upon price levels—Mises stressed the influence of people's preferences and expectations upon the demand for money, relative prices, and the structure of production. The volume on *Socialism* appeared in its first German edition in 1922 (first English edition in 1936). In this work Mises presented, besides a fascinating history and analysis of socialist thought, his own theory of economic calculation in a socialist economy. Socialist writers had never paid any attention to the problem of allocating productive resources efficiently, nor to the significance of pricing productive services, intermediate products, and final goods in a way that would assure consistency of economic planning with the preferences of the planning authorities, to say nothing about consistency with consumer preferences. Mises showed that competitive markets alone could generate the prices that would guide the allocation of resources in an efficient and consistent manner. The book *Human Action: A Treatise on Economics* appeared in English in 1949 (after a less elaborate German version had been published in 1940). This tome of 889 pages, different in conception, organization, epistemology, and ideology from any other treatise on economics, gave Mises an opportunity to compound his dissent from contemporary economists.

The profession reacted by withholding many of the honors customarily bestowed on its master builders. Neither in his twenty years at the University of Vienna nor in his twenty-five years at New York University was Mises appointed to a full chair, nor was he showered with a multitude of honorary degrees. Still, he was awarded honorary doctorates by New York University and the University of Freiburg; and in 1969, when he was 88 years old, the American Economic Association elected him a Distinguished Fellow.

◆

NOTES

Starting January 1, 1974, all manuscripts submitted to this *Review* have been sent out to referees without the author's name or affiliation. Original submissions should have a separate cover page for the title, author's name, and affiliation. The first page of the text should carry the title of the paper. For purposes of refereeing, authors are requested to avoid self-identification in the main text of their papers.

Nominations for AEA Officers

The Electoral College on March 8 chose Franco Modigliani as nominee for President-Elect of the American Economic Association in the balloting to be held in the autumn of 1974. Other nominees (chosen by the nominating committee) are: for Vice-President (two to be elected), Martin Bronfenbrenner, Leonid Hurwicz, James N. Morgan, and Charles L. Schultze; for member of the Executive Committee (two to be elected), Carolyn S. Bell, Jack Hirshleifer, Dale Jorgenson, and Burton A. Weisbrod.

Under a change in the bylaws as described in the *Papers and Proceedings* of this *Review*, May 1971, page 472, additional candidates may be nominated by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-Elect and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the Secretary at cost.

1975 Nominating Committee of the AEA

In accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended in 1972, President-Elect R. A. Gordon has appointed a Nominating Committee for 1975 consisting of Robert Eisner, Chairman; Bernard E. Anderson, Robert E. Gallman, Richard W. Parks, L. A. Rapping, Barbara B. Reagan, and Christopher A. Sims. Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-elect, the Committee shall include any other member of the Association nominated by petition including signatures and addresses of not less than 2 percent of the members of the Association, delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee. The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various offices to the Committee."

Annual Meeting Placement Service

The Placement Service at the 1974 annual meetings of the Allied Social Sciences Associations in San Francisco will begin operation on December 27, the day before sessions begin. Those who wish to do so will be able to attend more sessions with the first day set aside for labor market transactions. The Placement Service will continue during the other days of the annual meetings, but hopefully at a less frantic pace than in previous years.

Visiting Economics Scholars Program

At its meeting on December 27, 1973, the Executive Committee of the American Economic Association approved a proposal for a Visiting Economics Scholars Program by Edward J. Powers of Northern Michigan University. The program will be modeled on the Visiting Scientists Program formerly funded by the National Science Foundation and will begin with the academic year 1974–75. The goal is to provide visiting scholars to the more geographically isolated and less affluent colleges and universities across the country. For more information, contact C. Elton Hinshaw, Assistant Secretary-Treasurer, American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212.

Members with suggestions for the meetings in Dallas, Oct. 2–4, 1975 are invited to send them to Rendigs Fels, Secretary, American Economic Association, 1313—21st Avenue South, Nashville, Tennessee 37212. The Secretary will review the suggestions and send them to the Program Chairman. Economists wishing to give papers on econometrics or economic theory may want to submit abstracts to the Econometric Society, which meets with the Association and annually schedules a substantial number of contributed papers. Whether the Program Chairman of the American Economic Association will want to schedule sessions with contributed papers is not known at this time, but those wanting to give noneconometric papers are invited to send abstracts to the Secretary. To be considered, proposals for this Association must be received by the Secretary no later than September 15, 1974.

The World Congress of the International Economic Association, to be held in Budapest, Hungary from August 19 to 24, 1974, is open to all members of the American Economic Association. The program was published in the March issue of this *Review*, pp. 244–46. Inquiries and requests for registration forms and hotel reservations may be sent to the International Economic Association, 54 Boulevard Raspail, 75270 Paris, Cédex 06, France, or the Bureau of the Fourth World Congress of Economists, Kossuth Lajos ter. 11, 1860 Budapest 55, Hungary. The registration fee is Swiss francs 120 if paid before July 15, Sw. f. 150 if paid later. Payments can be made in all convertible currencies at the current exchange rates.

There will probably be a group flight to the 4th World Congress of the International Economic Association in Budapest. Tentative plans are for departure from New York on August 15, from Budapest on August 30, at a cost of $501. For details, Contact Mr. George Marucci, National Center for Educational Travel, 13003 Occoquan Road, Woodbridge, VA 22191 (703-491-1001 or 1009) before July 15. Limited to members of the American Economic Association of at least six months standing and their families.

The 14th World Conference of the Society for International Development will be held August 12–16, 1974. at the Hotel Ivoire in Abidjan, Ivory Coast. The conference theme is "Confrontation or Cooperation?" and will deal primarily with the position of the Third World vis-à-vis the critical global problems—food, energy, population, etc. For further information as to travel, hotel and conference registration, contact International Headquarters, Society for International Development, 1346 Connecticut Ave., N. W., Washington, D. C. 20036.

Maritime Studies and Management, an international journal, started publication in July 1973. The fourth issue appeared April 1974. This interdisciplinary academic and management oriented journal focuses on ocean and coast-related economic, legal, managerial, and social issues. Manuscripts are invited and may be sent to Professor Robert Vambery, Rutgers University, Graduate School of Business, Newark, New Jersey 07102.

The first annual convention of the Eastern Economic Association will be held Oct. 25–27, 1974 at the State University of New York at Albany, New York. Send papers, abstracts, inquiries to Dr. T. S. Saini, Bloomsburg State College, Bloomsburg, Pennsylvania 17815, telephone 717+389−2208.

Each year since 1954 the Council of the Operations Research Society of America has offered the Lanchester Prize for the best English-language published contribution in operations research. The prize for 1973 consists of $2,000 and a commemorative medallion. The screening of books and papers for the 1973 prize will be carried out by a committee appointed by the Council of the Society. To be eligible for consideration, the book or paper must be nominated to the committee. Nominations may be made by anyone. They should be sent to Philip M. Morse, Chairman, 1973 Lanchester Prize Committee, Room 24-115, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

A new edition of the volume, *Graduate Study in Economics: A Guide to Graduate Programs in Economics and Agricultural Economics at United States and Canadian Universities*, is being prepared in the offices of the Economics Institute, under the joint auspices of the American Economic Association and the American Agricultural Economics Association. The *Guide* was last revised in 1969; the new edition is to be considerably expanded. It should be available for purchase early in the fall. Order details will be carried in the September issue of this *Review*.

The 1974 Annual Meeting of the African Studies Association will strive to broaden the perspectives on Africa by focusing attention on African culture, society, economics, and politics as they may emerge in the next decade. The meeting will be held Oct. 30–Nov. 2, 1974. The program chairman is Sam C. Sarkesian, Department of Political Science, Loyola University, 6525 North Sheridan Road, Chicago, Illinois 60626.

Modern Greek Society is a newsletter published by and for historians and social scientists with a special interest in Greece. It also aims at promoting contacts and exchanges between scholars studying Greece, and those studying other societies of Mediterranean Europe, and hopes to contribute to the growing interest in comparative analysis of the area. Subscriptions are $2.00 per year, and it is scheduled to appear twice each academic year. Correspondence should be addressed to Modern Greek Society, P.O. Box 102, New Hampton, New York 10958.

The Latin American Studies Association wishes to draw attention to a resolution approved at the Fourth National Meeting held in Madison, Wisconsin, May 1973: *Resolution #5* Whereas the developing areas of the world are losing primary resource materials for research through purchase by private and public collectors and research institutions of foreign countries, it is moved that when papers, documents, books, and other materials of historical research value are purchased or otherwise removed from countries where their absence will seriously prejudice the development of local research, the Executive Council of the Latin American Studies Association (LASA) singly and if possible in conjunction with other relevant professional and scholarly associations, will try to assure that the purchaser will arrange to leave on deposit in an appropriate research facility of the country a usable copy of the materials. Furthermore, this motion shall be communicated to all relevant professional and scholarly associations and, insofar as practical, to collectors and their agents who might be involved in such transactions.

The new *Journal of Economic Criticism* will begin publication in the summer of 1974. The goals of this journal are to improve economic scholarship and to foster debate on the assumptions employed by scholars of business and economics. It will accomplish its goals through publication of comments and criticisms of scholarly work published elsewhere (including the *American Economic Review*). There will be numerous book reviews which will be designed to discuss the scholarship and assumptions of the author rather than an abstract of the book. A unique feature of this journal will be the regular publication of numerous short, perhaps only a paragraph in length, articles calling attention to and correcting errors detected in scholarly publications. To submit articles or inquire about subscriptions write to Roger K. Chisholm, Editor, Department of Economics, Memphis State University, Memphis, Tennessee, 38152.

Promotions

Dan M. Bechter: research officer and economist, Federal Reserve Bank of Kansas City, Jan. 2, 1974.

Sidney M. Blumner: professor, department of economics, California State Polytechnic University, Sept. 1, 1973.

C. Edward Harshbarger: research officer and economist, Federal Reserve Bank of Kansas City, Jan. 2, 1974.

James Heilbrun: professor of economics, Fordham University, Sept. 1974.

Robert S. Lande: associate professor, department of economics, Monterey Institute of Foreign Studies, Feb. 1, 1974.

Fred J. Levin: chief, Money and Finance Division, Federal Reserve Bank of New York, Jan. 24, 1974.

Charles M. Lucas: manager, statistics department, Federal Reserve Bank of New York, Jan. 1, 1974.

Anton S. Nissen: assistant vice president, Research and Statistics Function, Federal Reserve Bank of New York, Jan 1, 1974.

Brian Rungeling: associate professor, department of economics and finance, University of Mississippi, July 1, 1973.

Dominick Salvatore: associate professor of economics, Fordham University, Sept. 1974.

Robert Schafer: assistant professor, department of city and regional planning, Harvard University, July 1973.

Nancy Z. Spillman: assistant professor of economics, Los Angeles Trade-Technical College, Dec. 1973.

Ronald D. White: assistant professor of economics and director, Center for Business and Economic Research, Wichita State University, Jan. 7, 1974.

Administrative Appointments

John J. Casson, Brown Brothers Harriman & Co.: economist, American Express Company.

B. F. Kiker: head, department of economics, University of South Carolina, July 1, 1973.

Appointments

Jarvis M. Babcock: assistant professor, department of economics, Oberlin College.

Gilbert Bassett: assistant professor of economics, University of Illinois at Chicago Circle, Sept. 1973.

Martin Bronfenbrenner, Duke University: professor of economics, Wellesley College, fall semester 1974.

Antonio Camacho: professor of economics University of Illinois at Chicago Circle, Sept. 1973.

William L. Carlson: assistant professor of economics, St. Olaf College, 1973–74.

Jeffrey I. Chapman, Institute of Government and Public Affairs: assistant professor of public administration, University of Southern California, Sept. 1973.

Benjamin Chinitz, Brown University: professor of economics and director, Economic Growth Institute, State University of New York, Binghamton Feb. 1974.

Bradley T. Gale, University of Massachusetts: visiting research fellow in business administration, Harvard University, Sept. 1973.

Richard L. Haney, Jr., Indiana University: assistant professor, University of Georgia, Sept. 1, 1973.

Michael Jesse: economist, banking studies department, Federal Reserve Bank of New York, Aug. 14, 1973.

Elinda B. F. Kornblith, Rochester University: instructor, department of economics, Wellesley College, Sept. 1973.

Helen L. Ladd, Harvard University: assistant professor, department of economics, Wellesley College, Sept. 1974.

Harold S. Luft: assistant professor of health economics, Health Service Administration Program, Stanford University, Sept. 1973.

Alicia Munnell, Harvard University: lecturer, department of economics, Wellesley College, Feb. 1974.

Rutbert D. Reisch: economist, Foreign Research Division, Federal Reserve Bank of New York, Nov. 8, 1973.

Eunice H. Shields, University of Wisconsin: instructor, department of economics, Wellesley College, Sept. 1973.

Roger W. Spencer, Federal Reserve Bank of St. Louis: visiting professor, department of economics, University of South Carolina, Jan. 1, 1974.

Daniel J. Sullivan, University of Minnesota: instructor of economics, Middlebury College, Feb. 1974.

Thomas F. Tabasz, State University of New York, Binghamton: instructor of economics, Ohio Wesleyian University.

A. S. Vasconcellos, University of West Florida: visiting associate professor, department of economics, University of South Carolina, Jan. 1, 1974.

Leaves for Special Appointments

A. K. Barakeh, University of South Alabama: senior economist, Alabama Energy Management Board, Montgomery, Jan. 1, 1974.

Resignations

Josef C. Brada, Jr., Ohio State University: Graduate School of Business Administration, New York University, Sept. 1974.

P.A.V.B. Swamy, Ohio State University: Board of Governors, Federal Reserve System, Jan. 1974.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, John W.), his present place of employment or enrollment: his new title (if any), his next place of employment (if known or if changed), and the date at which the change will occur.

C. Type each item on a separate 3×5 card, and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, November 1; *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1974 issue of the *Review* will carry the seventy-first list of doctoral dissertations in political economy in American universities and colleges. The list will specify doctoral degrees conferred during the academic year terminating June 1974. This announcement is an invitation to send us information for the preparation of the list. This announcement supercedes and replaces a letter which was sent annually from the managing editor's office.

The *Review* will publish in its December 1974 issue the names of those who will have been awarded the doctoral degree since June 1973 and the titles of their dissertations. Dissertation abstracts will no longer be published, as these are published elsewhere.

By June 30, please send us this information on 3×5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

JEL Classification No. ________
Name: LAST NAME IN CAPS: First Name, Initial
Institution Granting Degree: ________________________
Degree Conferred (Ph.D. or D.B.A.) ____________ Year ____________
Dissertation Title: ________________________

When degrees in economics are awarded under different names, such as Business Administration, Public Administration, or Industrial Relations, candidates in these fields whose training has been *primarily in economics* should be included.

All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.